UFRRJ

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA, TECNOLOGIA E INOVAÇÃO EM AGROPECUÁRIA

TESE

Predição e Mapeamento de Propriedades de Solos no Parque Nacional de Itatiaia com Sensoriamento Remoto Proximal e Imagens Orbitais Hiperespectrais

Yuri Andrei Gelsleichter

2020

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA, TECNOLOGIA E INOVAÇÃO EM AGROPECUÁRIA

PREDIÇÃO E MAPEAMENTO DE PROPRIEDADES DE SOLOS NO PARQUE NACIONAL DE ITATIAIA COM SENSORIAMENTO REMOTO PROXIMAL E IMAGENS ORBITAIS HIPERESPECTRAIS

YURI ANDREI GELSLEICHTER

Sob orientação de
Lúcia Helena Cunha dos Anjos

e co-orientação de
Paula Debiasi

Tese submetida como requisito parcial para obtenção do grau de **Doutor** no Programa de Pós-graduação em Ciência, Tecnologia e Inovação em Agropecuária, Área de Concentração em Recursos Naturais e Proteção Ambiental.

Seropédica, RJ
Março de 2020

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

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

```
        Gelsleichter, Yuri Andrei, 1985-
G314p      Predição e mapeamento de propriedades de solos no
        Parque Nacional de Itatiaia com sensoriamento remoto
        proximal e imagens orbitais hiperespectrais / Yuri
        Andrei Gelsleichter. - Seropédica, 2020.
           89 f.: il.

           Orientadora: Lúcia Helena Cunha dos Anjos.
           Coorientadora: Paula  Debiasi.
           Tese(Doutorado). -- Universidade Federal Rural do
        Rio de Janeiro, Programa de Pós--Graduação em Ciência,
        Tecnologia e Inovação em Agropecuária, 2020.


           1. Mapeamento hiperespectral de solos. 2. Predição
        espectral de solos. 3. Covariáveis espectrais. 4.
        Sensoriamento remoto proximal de solos. I. dos Anjos,
        Lúcia Helena Cunha, 1957-, orient. II. Debiasi, Paula
        , 1982-, coorient. III Universidade Federal Rural do
        Rio de Janeiro. Programa de Pós--Graduação em Ciência,
        Tecnologia e Inovação em Agropecuária. IV. Título.
```

Este documento foi criado usando o sistema LaTeX de preparação de documentos para composição de alta qualidade originalmente desenvolvido por Leslie Lamport a partir do sistema de formatação TeX criado por Donald Knuth.
O formato final deste documento foi obtido usando uma derivação da classe `UFRuralRJ`, uma adaptação livre das classes `mdtufsm` e `iiufrgs` para a formatação de documentos acadêmicos produzidos na Universidade Federal Rural do Rio de Janeiro (UFRRJ) de acordo com as recomendações contidas na terceira edição do *Manual de instruções para organização e apresentação de dissertações e teses na UFRRJ*, publicado no ano de 2006.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) - Código de Financiamento 001. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA, TECNOLOGIA E INOVAÇÃO EM AGROPECUÁRIA

YURI ANDREI GELSLEICHTER

Tese submetida como requisito parcial para obtenção do grau de **Doutor** no Programa de Pós-graduação em Ciência, Tecnologia e Inovação em Agropecuária, Área de Concentração em Recursos Naturais e Proteção Ambiental.

TESE APROVADA EM 26/03/2020.

____
Lúcia Helena Cunha dos Anjos. Ph.D. UFRRJ
(Orientadora)

____
Mauro Antonio Homem Antunes. Ph.D. UFRRJ

____
Helena Saraiva Koenow Pinheiro. Dra. UFRRJ

____
Lênio Soares Galvão. Dr. INPE

____
Marcos Rafael Nanni. Dr. UEM

Dedico a minha família.

AGRADECIMENTOS

Agradeço a Deus pelo dom da vida, inteligência e pela capacidade de realização pelos quais sou agraciado. A minha família, que desde de criança, sempre deu exemplo, incentivou ao trabalho e estudo, as descobertas e a curiosidade, em especial meu pai Arlindo que sempre com muita paciência respondeu os meus infinitos questionamentos que despertaram ali o meu lado científico. A minha querida mãe que, igualmente, com seu exemplo de dedicação incansável e trabalho forte. A minha doce esposa Maria Aparecida Marques que sempre incentivou e encorajou todas as tomadas de decisões. Inclusive abdicando do que fosse preciso. E também sua família.

Agradeço a todos os professores que contribuíram para subir cada degrau desta escada da vida. E mais recentemente, a professora Lúcia Helena Cunha dos Anjos, que foi mais que um exemplo para mim, despertando o meu interesse cada vez mais em solo, principalmente, em gênese e classificação de solos. Com muita sensatez cobrando quando preciso e com grande paciência para ver os frutos deste trabalho florescerem, demonstrando, a cada dia, liderança e firmeza admiráveis. Por todos os e-mails, mensagens e respostas quase instantâneas. E ainda pelo curso que ministrei de *R* junto Com Shirlei e Gilsonley, e a oportunidade colaborar em programação em *R* com Catalina Bozzer, que agradeço também. A professora Erika Michéli da universidade Szent István da Hungria que me fez despertar para a ciência do solo sendo também uma pessoa direta e com natural liderança admiráveis. A essas duas professoras que perceberam em mim uma capacidade latente e deram-me a chance de demonstrar o meu potencial.

Também agradeço aos professores do Programa de Pós-graduação em Agronomia Ciência do Solo, principalmente a Helena Saraiva Koenow Pinheiro pela disciplina de Mapeamento Digital de Solos, que mudou a minha vida, fazendo com que eu me mudasse por completo o meu olhar sobre programação, e consequentemente fiquei fascinado com as possibilidades em Mapeamento Digital de Solos, algoritmos de aprendizado de máquinas e tratamento de dados. Igualmente aos pesquisadores da Embrapa Solos, Waldir de Carvalho Junior e Cesar da Silva Chagas que integraram a mesma disciplina e sempre estavam prontos para suprir qualquer dúvida. Ao Fabiano Baliero pelas considerações sobre solo e carbono.

Aos professores de graduação Elisa Helena Siegel Moecke pela iniciação científica e Gabriel Oscar Cremona Parma pelo despertar em Geoprocessamento.

O professor Mauro Antonio Homem Antunes que foi sempre amigo próximo, além das discussões técnicas e suas aulas muito detalhadas e informativas em Sensoriamento Remoto. Pelo apoio com equipamentos e materiais de laboratório. A professora Paula Debiasi, pela disposição em ser coorientadora, e sempre pronta a colaborar.

A Gerard B.M. Heuvelink e Rik van Heumen pelas sugestões de método.

Aos amigos e colegas do dia a dia, em destaque Elias Mendes Costa por transmitir a sua admiração pelo Software/linguagem de programação *R*, o que despertou-me para além das maneiras convencionais de processar, analisar, visualizar dados e a infinidades de operações em *R*, mas principalmente pela liderança e colaboração nas coletas a análises de solo do Parque Nacional de Itatiaia, juntamente com Robson Altiellys Tosta Marcondes, e toda equipe do LGCS-UFRRJ. Sobre visualização de dados em *R*, incluo aqui o Professor Márk Szalai da mesma universidade húngara. Ao Alessandro Samuel Rosa que muito inovou no modo pesquisar nos abrindo portas de possibilidades, para mim, principalmente demostrando o potencial do Software/linguagem de programação *R*. A Ana Paula Pessim de Oliveira pela amizade, suprimento de dúvidas e também pelo auxílio já na minha inscrição do curso. Aos coordenadores do Programa de Pós-Graduação em Ciência Tecnologia e Inovação em Agropecuária, bem como a Renata Gomes sempre pronta a amparar os alunos do curso com tudo que pode. A todos os demais por todas as discussões e contato.

Em relação ao experimento com as leituras espectrais de solos agradeço: Aos professores Lúcia e Mauro por toda orientação, assim como todos os contatos que proporcionaram acesso aos materiais e equipamentos; Gabriela Valente pelo apoio, empenho e dedicação; Professor Alexandre Demattê pela confirmação da metodologia; Ted King pela dedicação e empenho com o laboratório; Gabriela Lira pelo treinamento com o equipamento; Professora Maria Ivone pela disposição de laboratório e equipamentos; Professor Leonardo Duarte pela autorização de uso de equipamentos no Instituto de Tecnologia; Lucinano pelas orientações e explicações e dicas sobre secagem de materiais; Vinícius pelo apoio com os materiais e dissecadores; Flávio, Monalisa, Ednaldo e Claudinho pelos materiais emprestados e placas de petri da Embrapa Agrobiologia; Mauro (FAPUR) pelos trâmites com o equipamento; Maria pelo apoio incondicional sempre.

A CAPES pela concessão da bolsa de estudos, a FAPERJ pelo financiamento da pesquisa e Finep pela aquisição do espectrorradiômetro para o laboratório LabSpec. A TEMPUS FOUNDATION e CLIFF-GRADS pelas bolsas de intercâmbio. Os gestores do Parque Nacional de Itatiaia pela colaboração na pesquisa. O ICMBio pela autorização da pesquisa. O Ministério do Meio Ambiente pelas imagens de satélite RapidEye.

Aos desenvolvedores de softwares. Principalmente os de livre acesso e código aberto. Em especial para este trabalho *R*, Quantum GIS, SAGA GIS, e Linux.

Agradeço também a Universidade Federal Rural do Rio de Janeiro, juntamente com seus funcionários e colaboradores. E agradeço aos que, por meio de impostos financiam: a universidade, sua operação e manutenção, seus equipamentos, o experimento e também minha bolsa de estudos que foi muito importante para a realização deste trabalho.

Aos professores Doutores que participaram de minha banca, agradeço por todos os questionamentos, sugestões e correções.

Aos que não estão aqui, e de qualquer modo estiveram presentes ou contribuíram indiretamente, também agradeço. Todos foram importantes para o trabalho. Mais uma vez, deixo aqui o meu muito obrigado!

RESUMO GERAL

GELSLEICHTER, Yuri Andrei. **Predição e mapeamento de propriedades de solos no Parque Nacional de Itatiaia com sensoriamento remoto proximal e imagens orbitais hiperespectrais**. 2020. 89f. Tese (Doutorado em Ciência, Tecnologia e Inovação em Agropecuária). Pró-Reitoria de Pesquisa e Pós-Graduação, Universidade Federal Rural do Rio de Janeiro, Seropédica, RJ, 2020.

O Parque Nacional de Itatiaia (INP, do inglês para *Itatiaia National Park*) fica localizado ao sul do estado do Rio de Janeiro, com o estudo realizado na Parte Alta do Parque, definida acima de 2000 msnm. Os objetivos do primeiro capítulo foram: (i) investigar a capacidade de predizer propriedades do solo (Al, Ca, K, Mg, Na, P, pH, Carbono Total (TC *Total Carbon*), H e N), utilizando os comprimentos de onda 350–2500 nm; e (ii) investigar e desenvolver pré-processamentos espectrais para uso e comparação em algoritmos de aprendizado de máquina, como Redes Neurais Artificiais (ANN, *Artificial Neural Networks*), *Random Forest* (RF), Regressão de Mínimos Quadrados Parciais (PLSR, *Partial Least Squares Regression*) e *Cubist* (CB). Foram coletadas amostras de solo, por horizontes, em 84 perfis de solo, compondo um total de 300 amostras. A validação cruzada aplicada para avaliar os modelos foi do tipo k-fold. O melhor pré-processamento espectral foi o Inverso da Reflectância de Fator 10^4 (IRF4) para TC com CB que superou os métodos comumente utilizados, com coeficiente de determinação (R^2) médio de 0,85, RMSE de 1,96 para TC; e 0,67 com 0,041, respectivamente, para H. Para o mapeamento do TC nos solos do INP foram utilizadas três cenas de imagens hiperespectrais do sensor *Compact High Resolution Imager* (CHRIS) do satélite (plataforma espacial) *Project for On Board Autonomy* (PROBA). Este sensor conta com 62 bandas espectrais no intervalo dos comprimentos de onda 406 a 1019 nm (referente as bordas da primeira e última bandas respectivamente). As imagens foram corrigidas quanto a ruídos, *striping*, distorções geométricas e interferências atmosféricas. A predição de TC foi feita usando essas imagens e associando covariáveis de relevo e imagens do sensor orbital RapidEye, obtendo R^2 de 0,33. Utilizando-se apenas a cena RapidEye mais as covariáveis de terreno o R^2 foi de 0,32. Essas imagens foram combinadas aos espectros proximais obtidos na primeira camada do solo, dos 84 perfis, para produzir imagens de refletância de solo de toda parte alta do INP. Essa técnica foi chamada de imageamento espectral de subsuperfície. A aplicação deste produto no Mapeamento Digital de Solos aumentou significativamente a predição de TC, com R^2 de 0,58, com incremento de 75% em relação ao Mapeamento Digital de Solos convencional. Essa técnica inovadora, apresentada pela primeira vez neste estudo, é denominada Mapeamento Hiperespectral de Solos (HSM, em inglês *Hyperspecrtal Soil Mapping*), sendo o desenvolvimento desta técnica o objetivo principal do segundo capítulo. Essa técnica pode isolar o efeito de interferência atmosférica e efeitos de cobertura de solo e vegetação sobre a reflectância do solo. Pelo aumento da capacidade de predição do HSM, pode-se reduzir a quantidade amostral do levantamento pedológico, alcançando assim resultado equivalente ao Mapeamento Digital do Solos. O HSM é ideal para áreas com acesso e locomoção muito restritos, como o INP, mas também pode ser aplicado para o mapeamento de atributos de solo, fins agrícolas e monitoramento ambiental.

Palavras-chave: Covariáveis espectrais. Predição espectral de solos. Mapa hiperespectral.

GENERAL ABSTRACT

GELSLEICHTER, Yuri Andrei. **Predicting and mapping soil properties in Itatiaia National Park with proximal remote sensing and hyperspectral orbital images**. 2020. 89p. Thesis (Doctorate in Science, Technology and Innovation in Agriculture). Pró-Reitoria de Pesquisa e Pós-Graduação, Universidade Federal Rural do Rio de Janeiro, Seropédica, RJ, 2020.

The Itatiaia National Park (INP) is located Southern of Rio de Janeiro State, in the boundary with Minas Gerais and São Paulo states, Southeast region of Brazil. This study was carried out in the Upper Part of the INP, defined above the 2000 msnm. The objectives of the first chapter of this study were: (i) to investigate the ability to predict soil properties (Al, Ca, K, Mg, Na, P, pH, Total Carbon (TC), H and N), using wavelengths 350–2500 nm; and (ii) to investigate and develop spectral preprocessing for usage and comparison in machine learning algorithms, such as Artificial Neural Networks (ANN), Random Forest (RF), Partial Least Squares Regression (PLSR) and Cubist (CB). In the Upper Part of the INP soil samples were collected from the horizons of 84 soil profiles, composing a total of 300 samples. The cross-validation method used to evaluate the models was the k-fold type. The best spectral preprocessing was the Inverse of Reflectance to Factor of 10^4 (IRF4) for TC with CB. IRF4 surpassed the common methods used for preprocessing, with an average coefficient of determination (R^2) of 0.85, RMSE of 1.96 for TC; and 0.67 with 0.041, respectively, for H. The results pointed out IRF4 as one of the best preprocessing associated with the RF and CB algorithms. To map the TC in the INP soils, there were used three scenes of Hyperspectral images from the Compact High-Resolution Imager (CHRIS) sensor from space platform Project for On Board Autonomy (PROBA), a satellite of the European Spatial Agency (ESA). This sensor contains 62 spectral bands in the wavelengths interval of 406 to 1019 nm (as reference, the edge of the first and last bands respectively). The images were corrected for noise, striping, geometric distortions and atmospheric interferences. The TC prediction was made using these images and associating relief covariates and images from the RapidEye orbital sensor, obtaining R^2 of 0.33. Using only the RapidEye scene plus the terrain covariates the R^2 was 0.32. These images were combined with the proximal spectra obtained in the top soil layer, of the 84 profiles, to produce soil reflectance images of INP Upper Part. This technique was called Subsurface spectral imaging, with the application of this product in Digital Soil Mapping the TC prediction increased significantly, with R^2 0.58, showing an increase of 75% in relation to the conventional Digital Soil Mapping. This innovative technique is presented for the first time in this study, and is called Hyperspectral Soil Mapping (HSM). The development of this technique was the main objective of the second chapter. The spectral preprocessing image (in HSM) can isolate the effect of atmospheric interference and effects of the land cover and vegetation on the soil reflectance. Thus, by increasing the predictive capacity of the HSM, the sample size of the pedological survey can be reduced, having a result equivalent to the Digital Soil Mapping. In addition to reducing the cost of taking samples, this technique is ideal for areas with very restricted access and locomotion, as the case of INP, but it can also be applied for mapping of various soil properties, agricultural purposes and remote environmental monitoring.

Keywords: Spectral covariates. Spectral prediction of soils. Hyperspectral map.

RESUMEN AMPLIADO

GELSLEICHTER, Yuri Andrei. **Predicción y cartografía de las propiedades del suelo en el Parque Nacional de Itatiaia con imágenes orbitales hiperespectrales y de teledetección proximal**. 2020. 89p. Tesis (Doctorado en Ciencia, Tecnología e Innovación en Agricultura). Pró-Reitoria de Pesquisa e Pós-Graduação, Universidade Federal Rural do Rio de Janeiro, Seropédica, RJ, 2020.

1 Introducción

Las técnicas para cuantificar las propiedades químicas del suelo se basan en la titulación (química húmeda). El desarrollo de sensores y procesos computacionales para respaldar métodos más rápidos y limpios está en desarrollo continuo. Las técnicas de Teledetección (RS) como la Detección Proximal del Suelo (PSS) actualmente son aplicables a ciertas propiedades. Los métodos de procesamiento de datos espectrales permiten aumentar la capacidad de predecir las propiedades del suelo. Existe una amplia gama de algoritmos para probar y pueden funcionar de manera diferente según el preprocesamiento espectral.

En términos de Mapeo Digital de Suelos (DSM), hay una falta de uso/integración con PSS. En el segundo capítulo, exploramos la integración de PSS para DSM utilizando una técnica que denominamos Imagen de subsuperficie para promover lo que denominamos Mapeo de suelos hiperespectrales.

2 Material y Métodos

Este estudio se realizó en la Parte Alta dell Parque Nacional de Itatiaia (INP), que se encuentra por encima de los 2000 msnm. Es una región montañosa ubicada al sur del estado de Río de Janeiro, en el límite con los estados de Minas Gerais y São Paulo. Debido al reducido impacto de la agricultura de las últimas décadas, y como unidad de conservación, el INP es un área de referencia para los estudios ambientales.

Capítulo I

Se aplicó el RS como Proximal Soil Sensing (PSS) (350 a 2500 nm) para explorar los tratamientos espectrales para predecir las seguintes propiedades del suelo: aluminio (Al), calcio (Ca), potasio (K), magnesio (Mg), sodio (Na), fósforo (P), más pH, carbono total (TC), hidrógeno (H) y nitrógeno (N), con los algoritmos: Artificial Neural Network (ANN), Random Forest (RF), Partial Least Squares Regression (PLSR) y Cubist (CB).

Los principales preprocesos espectrales son Continuum Removal (CR) (CLARK, 1999), Savitzky–Golay (SVG) (SAVITZKY; GOLAY, 1964) con diferentes configuraciones en la derivada, orden del polinomio y la ventana de búsqueda (VASQUES *et al.*, 2008) e Inversa de reflectancia para Factor de 10^4 (IRF4). El IRF4 se obtuvo dividiendo 10.000 por cada valor del espectro de reflectancia. También se incluyó como preprocesamiento una conversión de datos espectrales a absorbancia mediante -log10 (reflectancia) (ROSSEL *et al.*, 2005) (AB-log). La forma de las principales curvas de preprocesamiento se puede observar en la Figura 1.

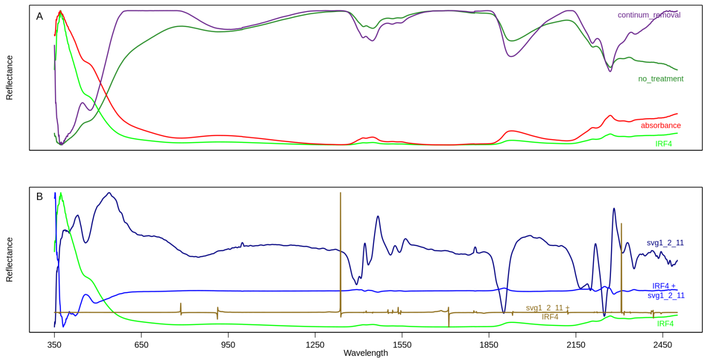


Figura 1: Visualización de las principales curvas de preprocesamiento espectral: (A) Eliminación de continuo (purpura); sin tratamiento (espectro sin procesar) (verde); absorbancia (rojo); Inverso de la reflectancia al factor de 10^4 (verde claro). (B) primera derivada de Savitzky-Golay (azul oscuro); Inversa de la reflectancia al factor de 10^4 + primera derivada de Savitzky-Golay (azul claro); Primera derivada de Savitzky-Golay + Inverso de reflectancia al factor de 10^4 (marrón); Inverso de la reflectancia al factor de 10^4 (verde claro). Observe que cada curva se ajusta a su propia escala *y* (reflectancia).

Capítulo II

Para saber si la cantidad puntos en el INP estaba suficiente se analizó la dependencia espacial con un semivariograma (Figura 2).

Para realizar la integración entre DSM y PSS utilizamos los valores de PSS del sensor ASD Fieldspec 4 (de muestras de suelo superior), imágenes del sensor Multiespectral RapidEye con 5 bandas y el Hyperspectral one CHRIS con 62 bandas. La imagen de CHRIS tuvo que ajustarse cuidadosamente para poder usarla. Los tratamientos consistieron en una corrección geométrica y una atmosférica, ajuste en tamaño de píxel, en intensidad de reflectancia entre las diferentes imágenes y tratamiento de sombras.

La integración de DSM y PSS se calcula con el enfoque DSM, pero en lugar de predecir una propiedad del suelo, se predice un valor de una banda de PSS. La salida es una imagen de una banda de PSS, se repitió el proceso llamado Imagen de Subsuperficie para generar 100 imágenes a partir de las longitudes de onda de PSS elegidas. La imagen de subsuperficie (Figura 3) se ejecuta el proceso de mapeo digital, llamado Mapeo Hiperespectral del Suelo (Figura 4).

La integración de DSM y PSS se calcula con el enfoque DSM, pero en lugar de predecir una propiedad del suelo, predice un valor de una banda de PSS.

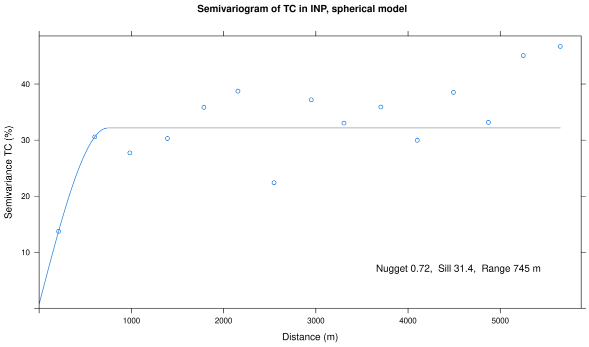


Figura 2: Semivariograma de TC en INP con modelo esférico.

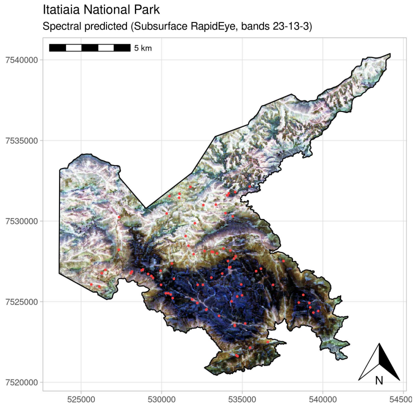


Figura 3: Imagen de subsuperficie de la parte Alta del INP.

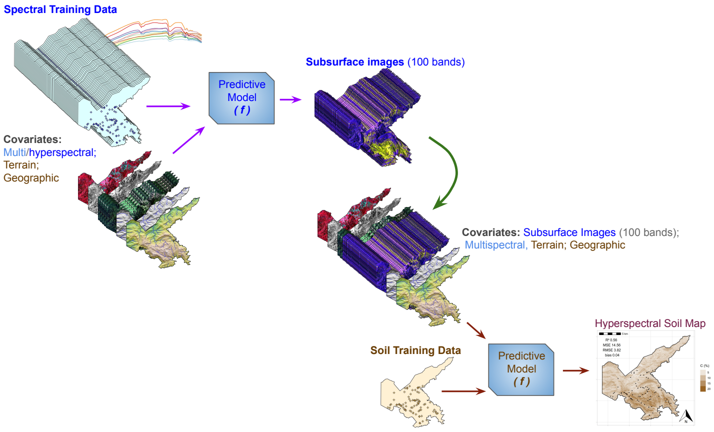


Figura 4: Proceso de subsuperficie, los procesos de Mapeo Hiperespectral de Suelos (HSM).

3 Resultados

Capítulo I

Para evaluar el desempeño de los modelos predictivos, se calcularan el error cuadrático medio (RMSE), el coeficiente de determinación (R^2) y la relación de desempeño a desviación (RPD). El mejor modelo asociado con el mejor preprocesamiento, fue el CB para TC con R^2 de 0,85, RPD de 2,87 (el más alto), seguido por PLSR para N con R^2 de 0,82 y RPD de 2,65, y RF para Al con R^2 de 0,54 y RPD de 1,54 (Tabla 1 y Figura 5).

Tabla 1: Los mejores preprocesamients con los modelos asociados para cada propiedad de los suelos muestreados en el INP.

Preprocessing	Model	Soil property*	R^2	MSE	RMSE	bias	RPD
IRF4 + SVG-1-2-11 + NR 434	rf	Al	0.536	0.944	0.954	0.037	1.541
IRF4	cb	H	0.672	0.173	0.411	-0.034	1.817
SVG-1-2-11 + IRF4 + NR 434	rf	K	0.275	0.017	0.118	0.003	1.244
SVG-1-2-11	rf	Mg	0.194	0.074	0.267	0.014	1.148
IRF4 + NR 434	plsr	N	0.819	0.018	0.13	-0.005	2.649
SVG-1-2-11 + IRF4 + NR 434	rf	P	0.072	66.896	7.436	0.137	1.07
SVG-1-2-11	rf	pH	0.363	0.096	0.309	-0.005	1.286
IRF4	cb	TC	0.852	3.998	1.958	-0.044	2.867

Ca y Na no se muestran en la tabla debido a no satisfactorios resultados. La descripción de cada preprocesamiento está de acuerdo con Tabla 2.1; rf: bosque aleatorio; cb: cubista; plsr: Regresión de mínimos cuadrados parciales; TC: carbono total; R^2: coeficiente de determinación; MSE: error cuadrático medio; RMSE: raíz del error cuadrada medio; RPD: relación entre rendimiento y desviación. Las unidades de los coeficientes corresponden a Tabla* 2.2. (*Las Tablas mencionadas están en el texto principal).

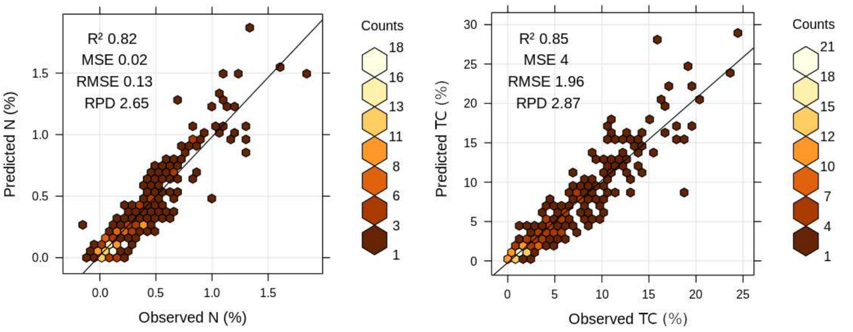


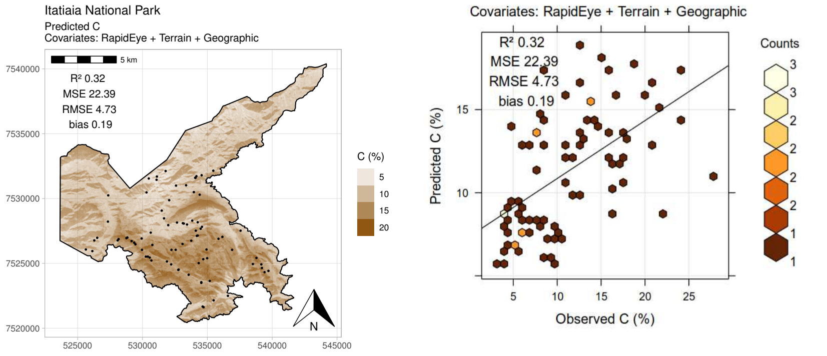
Figura 5: Relación entre los valores observados y predichos de N y TC.

El DSM que usa RapidEye alcanza un R^2 de 0,32 mientras que el HSM con la técnica de imagen del subsuperficie, logra un R^2 de 0,56, lo que proporciona un incremento del 75% en la capacidad de predicción (Table 2). Estos resultados pueden ser apreciados en los mapas de las Figuras 6a y 7a, en el primero de ellos, se observa una sobre estimación de los valores de TC, mientras que, en el segundo de ellos, la predicción lograda mostró una mayor exactitud como se visualiza en la Figura 7b. El rendimiento de CHRIS estuvo cerca del RapidEye con un ligero incremento en la capacidad de predicción.

Tabla 2: Estadística descriptiva de la predicción espacial del TC en el INP sobre las covariables utilizadas y la principal covariable explicativa.

Covariates groups	R^2	MSE	RMSE	bias	Main cov.
RapidEye + Terrain + Geographic (Without CHRIS)	0.32	22.39	4.73	0.19	DEM
Subsurface CHRIS (bands 1:100) + RapidEye + Terrain + Geographic	0.57	14.13	3.76	0.09	band 53
Subsurface RapidEye (bands 1:100)+RapidEye+ Terrain+ Geographic	0.56	14.56	3.82	0.04	band 75

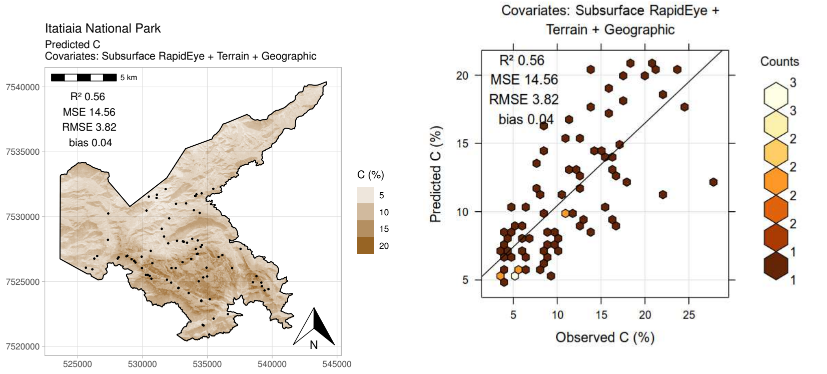
R^2: coeficiente de determinación; MSE: Error cuadrático medio en (%) como TC; RMSE: raíz del error cuadrada medio en (%) como TC; Covariable principal: covariable de mayor rango del correspondiente modelo de bosque aleatorio. Banda 53 = 895 nm y banda 75 = 1480 nm.



(a) Mapa TC RapidEye

(b) TC RapidEye, observados versus predichos

Figura 6: Predicción espacial del TC sobre en el INP, con las covariables RapidEye, Terrain, Geographic.



(a) Map TC Subsurface RapidEye 1 a 100 bandas

(b) TC Subsurface RapidEye 1 a 100 bandas, observados versus predichos

Figura 7: Predicción espacial del TC sobre en el INP, con las covariables Subsurface RapidEye (bandas 1 a 100), RapidEye (bandas multiespectrales 1 a 5), Terrain, Geographic.

4 Conclusións

Considerando el carbono del suelo como indicador de la salud, calidad y degradación del suelo, los resultados obtenidos a través de las técnicas de predicción de propiedades espectrales del suelo aplicadas muestran un gran potencial para una rápida evaluación ambiental. En este sentido, estas técnicas pueden contribuir a la gestión y seguimiento de la Parte Alta del Parque Nacional de Itatiaia.

Capítulo I

Cada propiedad del suelo tiene el potencial de predicción incrementado por un preprocesamiento espectral específico. Para TC, el IRF4 superó al preprocesamiento de uso común, incluido SVG. Para algunos casos, como en la aplicación de CR para predecir TC e IRF4 (solo) e para predecir N, el preprocesamiento disminuyó el potencial de predicción en comparación con los espectros no tratados. El algoritmo más presente entre los valores pronosticados más altos fue RF (5 de 8). La técnica IRF4 se introduce por primera vez en espectroscopia. Se recomiendan más estudios para confirmar el potencial de la herramienta de preprocesamiento.

Capítulo II

Por el análisis del semivariograma se entiende que el empleo de 84 puntos de muestreo de perfiles de suelo en el INP fue suficiente para realizar estudios de mapeo en el área. La combinación de PSS y DSM mediante el uso de imágenes subsuperficiales en HSM ha demostrado ser muy eficiente para mapear las propiedades del suelo, en comparación con el proceso DSM convencional. El HSM es la primera integración directa entre PSS y HSM.

Palabras clave: Covariables espectrales. Predicción espectral de suelos. Mapa hiperespectral.

EXTENDED ABSTRACT

GELSLEICHTER, Yuri Andrei. **Predicting and mapping soil properties in Itatiaia National Park with proximal remote sensing and hyperspectral orbital images**. 2020. 89p. Thesis (Doctorate in Science, Technology and Innovation in Agriculture). Pró-Reitoria de Pesquisa e Pós-Graduação, Universidade Federal Rural do Rio de Janeiro, Seropédica, RJ, 2020.

1 Introduction

The laboratory techniques to quantify chemical soil properties are normally based in the titration (wet chemistry). The development of sensors and computational processes to support a faster and cleaner methods are an ongoing development. Remote Sensing (RS) techniques such as the Proximal Soil Sensing (PSS) are already applicable to certain soil properties, and methods for preprocessing the spectral data can increase the capacity to predict these properties. There is a large range of algorithms to test and they can perform different according to the spectral preprocessing.

In terms of Digital Soil Mapping (DSM), there is a lack of use/integration with PSS. In the second chapter we explore the integration of PSS for DSM using a technique that we named as subsurfacing image to obtain what was named as Hyperspectral Soil Mapping.

2 Material and Methods

This study was carried out in the Upper Part of the Itatiaia National Park (INP), which is defined by the region above the 2000 msnm. The INP has a mountainous relief and is located Southern of Rio de Janeiro State, at the boundary with Minas Gerais and São Paulo states. The INP was the first national park, created in 1937, and since it a conservation unit, it is a reference area for environmental studies.

Chapter I

The RS as Proximal Soil Sensing (PSS) (350 to 2500 nm) was applied to explore the spectral preprocessing to predict the soil properties: aluminum (Al), calcium (Ca), potassium (K), magnesium (Mg), sodium (Na), phosphorus (P), plus pH, total carbon (TC), hydrogen (H) and nitrogen (N). There were used the algorithms: Artificial Neural Network (ANN), Random Forest (RF), Partial Least Squares Regression (PLSR) and Cubist (CB).

The main spectral preprocessing are Continuum Removal (CR) (CLARK, 1999), Savitzky-Golay (SVG) (SAVITZKY; GOLAY, 1964) with different settings across the derivative, order polynomial and search window (VASQUES *et al.*, 2008), and Inverse of Reflectance to Factor of 10^4 (IRF4). The IRF4 was obtained dividing 10,000 for each value of the reflectance spectrum. A conversion of spectral data to absorbance by the -log10 (reflectance) (ROSSEL *et al.*, 2005) (AB-log) was also included as a preprocessing. The main preprocessing curves' shape can be observed in Figure 1.

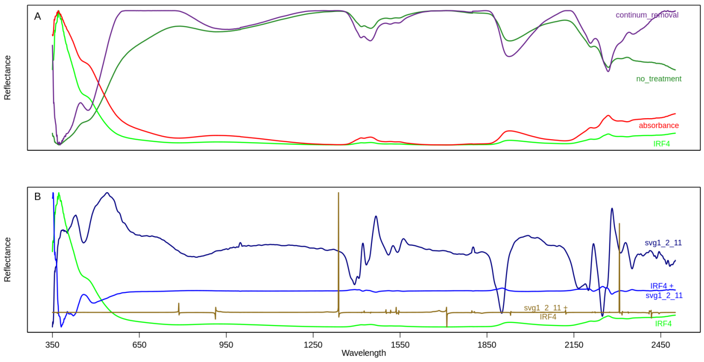


Figure 1: Visualization of main Spectral preprocessing curves: (A) Continuum Removal (magenta); no treatment (raw spectrum) (green); absorbance (red); Inverse of Reflectance to Factor of 10^4 (light green). (B) Savitzky-Golay first derivative (dark blue); Inverse of Reflectance to Factor of 10^4 + Savitzky-Golay first derivative (light blue); Savitzky-Golay first derivative + Inverse of Reflectance to Factor of 10^4 (brown); Inverse of Reflectance to Factor of 10^4 (light green). Notice, each curve fits its own *y* (reflectance) scale.

Chapter II

To find out if the number of points in the INP was sufficient, the spatial dependence was analyzed with a semivariogram (Figure 2).

To perform the integration between DSM and PSS we used the PSS values from sensor ASD Fieldspec 4 (from top soil samples), images from Multispectral sensor RapidEye with 5 bands and the Hyperspectral one CHRIS with 62 bands. The CHRIS image had to be carefully adjusted in order to be used. The treatments involved a geometric and atmospheric correction, adjust in pixel size, in reflectance intensity among the different images, and shadow treatment.

The DSM and PSS integration were calculated with DSM approach, but instead to predict a soil property it predicts a value of a band from PSS. The output is an image of a PSS band, the process was identified here as subsurfacing image, and it was repeated to generate 100 images from chosen PSS wavelengths. The subsurface image (Figure 3) was used to run the computational mapping process, named in this study as Hyperspectral Soil Mapping (Figure 4).

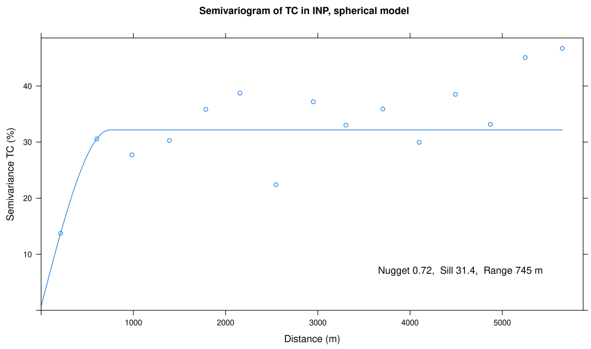


Figure 2: Semivariogram of TC in INP with spherical model.

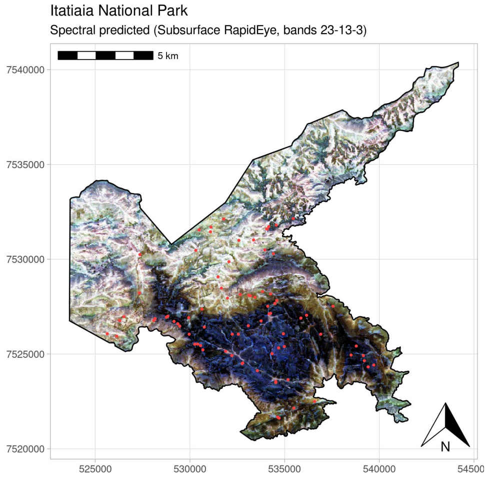


Figure 3: The upper part of INP Subsurface RapidEye image.

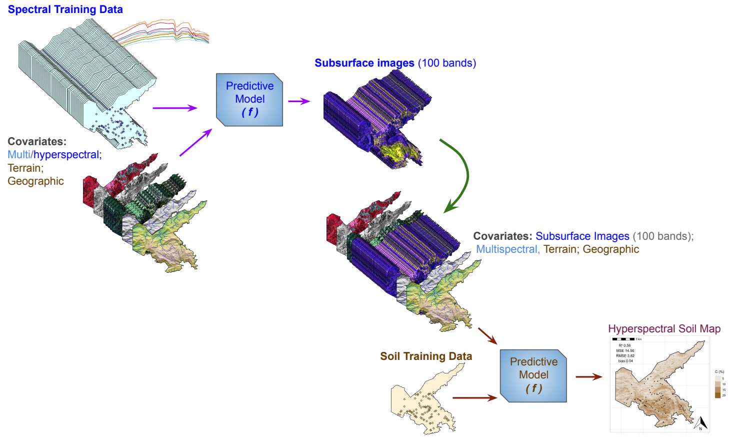

Spectral Training Data
Predictive Model (f)
Subsurface images (100 bands)
Covariates: Multi/hyperspectral; Terrain; Geographic
Covariates: Subsurface Images (100 bands); Multispectral, Terrain; Geographic
Hyperspectral Soil Map
Soil Training Data
Predictive Model (f)

Figure 4: Subsurfacing process, the steps to the Hyperspectral Soil Mapping (HSM) processes.

3 Results

Chapter I

To evaluate the performance of predictive models, the Root Mean Squared Error (RMSE), coefficient of determination (R^2) and Ratio of Performance to Deviation (RPD) were calculated. The best model associated with the best preprocessing, was the CB for TC with R^2 of 0.85, RPD of 2.87 (highest), followed by PLSR for N with R^2 of 0.82 and RPD of 2.65, and RF for Al with R^2 of 0.54 and RPD of 1.54 (Table 3 and Figure 5).

Table 3: Outstanding preprocessing with the associated models for each property of soils sampled at INP.

Preprocessing	Model	Soil property	R^2	MSE	RMSE	bias	RPD
IRF4 + SVG-1-2-11 + NR 434	rf	Al	0.536	0.944	0.954	0.037	1.541
IRF4	cb	H	0.672	0.173	0.411	-0.034	1.817
SVG-1-2-11 + IRF4 + NR 434	rf	K	0.275	0.017	0.118	0.003	1.244
SVG-1-2-11	rf	Mg	0.194	0.074	0.267	0.014	1.148
IRF4 + NR 434	plsr	N	0.819	0.018	0.13	-0.005	2.649
SVG-1-2-11 + IRF4 + NR 434	rf	P	0.072	66.896	7.436	0.137	1.07
SVG-1-2-11	rf	pH	0.363	0.096	0.309	-0.005	1.286
IRF4	cb	TC	0.852	3.998	1.958	-0.044	2.867

Ca and Na are not shown in the table due to the very poor results. The description of each preprocessing is according to Table 2.1; rf: random forest; cb: cubist; plsr: Partial Least Squares Regression; TC: total carbon; R^2: coefficient of determination; MSE: mean squared error; RMSE: root-mean-square error; RPD: ratio of performance to deviation. The coefficients units correspond to Table* 2.2. (*Mentioned Tables are in the main text).

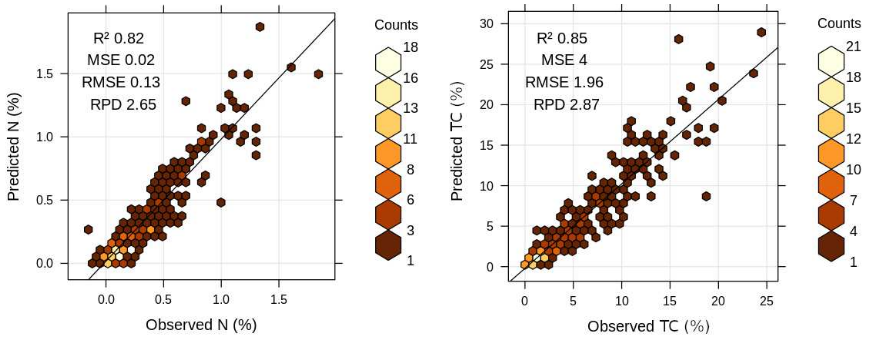


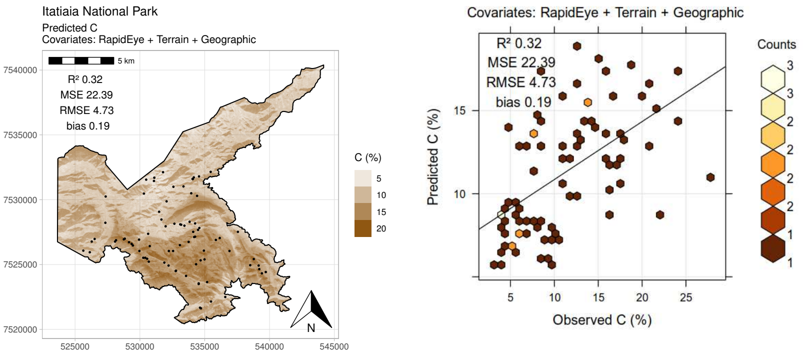
Figure 5: Relationship between the observed and predicted values of N and TC.

The DSM using RapidEye reaches an R^2 of 0.32 while using the HSM with Subsurface image the R^2 jump to 0.56, providing an increment gain of 75% in prediction capacity (Table 4). It is also possible to see the result in the maps on Figures 6a and 7a, the first shows an over estimation of TC while the second is more balanced, which is confirmed by the Figure 7b. The performance of CHRIS was close to that of RapidEye, with a slight increment of prediction capacity.

Table 4: Descriptive statistics of spatial prediction of TC on INP over the used covariates, and the main explanatory covariate.

Covariates groups	R^2	MSE	RMSE	bias	Main cov.
RapidEye + Terrain + Geographic (Without CHRIS)	0.32	22.39	4.73	0.19	DEM
Subsurface CHRIS (bands 1:100) + RapidEye + Terrain + Geographic	0.57	14.13	3.76	0.09	band 53
Subsurface RapidEye (bands 1:100)+RapidEye+ Terrain+ Geographic	0.56	14.56	3.82	0.04	band 75

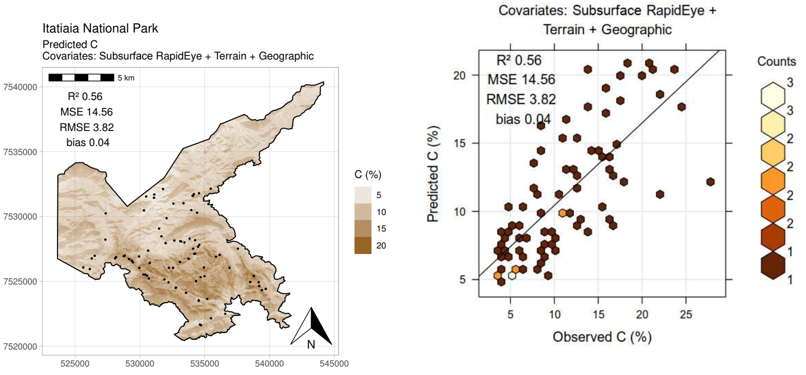
R^2: coefficient of determination; MSE: Mean Squared Error in (%) as TC; RMSE: Root Mean Square Error in (%) as TC; Main cov.: Higher ranked covariate from correspondent Random Forest model. Band 53 = 895 nm and band 75 = 1480 nm.



(a) Map TC RapidEye

(b) TC RapidEye, observed versus predicted

Figure 6: Spatial prediction of TC over INP, with the covariates RapidEye, Terrain, Geographic.



(a) Map TC Subsurface RapidEye 1 to 100 bands

(b) TC Subsurface RapidEye 1 to 100 bands, observed versus predicted

Figure 7: Spatial prediction of TC over INP, with the covariates Subsurface RapidEye (bands 1 to 100), RapidEye (multispectral bands 1 to 5), Terrain, Geographic.

4 Conclusions

Considering that soil carbon is an indicator of soil health, quality and degradation, the results obtained from the applied spectral soil properties prediction techniques show potential for fast environmental assessment. In this sense those techniques can contribute for the Itatiaia National Park management and monitoring.

Chapter I

Each soil property has the prediction potential increased by a specific spectral preprocessing. For TC, the IRF4 outperformed the commonly used preprocessing, including SVG. For some preprocessing of soil properties, such as CR for TC and IRF4 (alone) for N, the preprocessing decreased the potential for prediction in comparison with the non-treated spectra. The algorithm that was more frequently among the higher predicted values was the RF (5 out of 8). The IRF4 technique is introduced in spectroscopy here for the first time, thus, it is recommended more studies to confirm the potential as a preprocessing tool.

Chapter II

By the analysis of the semivariogram it is understood that the sampling mesh of 84 soil profiles was sufficient for sampling points in the mapping of the INP upper region. The combination of PSS and DSM through the use of subsurface image in the HSM was very efficient to map the soil properties, when compared with the normal DSM process. The HSM is the first direct integration between PSS and HSM.

Keywords: Spectral covariates. Spectral prediction of soils. Hyperspectral map.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIAÇÕES E SIGLAS

Algoritmos e softwares

ANN	Redes Neurais Artificiais (do inglês *Artificial Neural Network*)
CB	*Cubist*
cHLS	*Conditioned Latin Hypercube Sampling algorithm*
ML	Aprendizado de Máquina (do inglês *Machine Learning*)
R	Linguagem de programação
RF	*Random Forest*
PLSR	Regressão de Mínimos Quadrados Parciais (PLSR, do inglês *Partial Least Squares Regression*)
PSS	Sensoriamento Remoto Proximal do Solo (do inglês *Proximal Soil Sensing*)
SAGA-GIS	*System for Automated Geoscientific Analyses - Geographic Information System*
QGIS	*Quantum Geographic Information System*

Covariáveis

Aspect	*Represents exposure faces, values in degrees (0 to 360°)*
Convergence	*The general shape of the hillside in all directions (concave, rectilinear or convex)*
Cat_area	*Related to volume of flooding that reaches a certain cell*
CHNB	*Interpolation of a channel network base level elevation*
CHND	*Altitude above the channel network (CHNB - original elevation)*
DEM	Modelo Digital de Elevação (do inglês *Digital elevation model*)
LS_factor	*Attribute equivalent to the topographic factor of the Revised Universal Soil Loss Equation (RUSLE)*
NDVI	Índice de Vegetação por Diferença Normalizada (do inglês *Normalized Difference Vegetation Index*)
Northernness	*Indicates the direction of the slope relative to the north. Northernness = abs(180° −Aspect)*
Plan_curv	*The shape of the hillside on the horizontal plane (concave, rectilinear or convex)*
Prof_curv	*The shape of the hillside on the vertical plane (concave, rectilinear or convex)*
RSP	*Represents relative slope position based on the base channel network*
SAVI	Índice de Vegetação Ajustado ao Solo (do inglês *Soil-Adjusted Vegetation Index*)
Slope	*Gradient or rate of change of elevation between neighboring cells*
TWI	*Describes a tendency for a cell to accumulate water*

Distância

nm	Nanômetros
cm	Centímetros
m	Metro
msnm	Metros sobre nível do mar
Km	Quilômetros

Espectral

MIR	Comprimentos de onda do Infravermelho Médio (do inglês *Middle Infra-Red*)
NIR	Comprimentos de onda do Infravermelho Próximo (do inglês *Near Infra-Red*)
SWIR	Comprimentos de onda do Infravermelho de Ondas Curtas (do inglês *Short-Wave Infra-Red*)
UV	Comprimentos de onda do Ultravioleta
Vis	Comprimentos de onda do Visível – parte do espectro eletromagnético (luz) detectável pelos olhos humanos
Vis-NIR	Comprimentos de onda do Visível ao Infravermelho Próximo (do inglês *visible to near-infrared*)
Vis-NIR-SWIR	Comprimentos de onda do Visível ao Infravermelho de Ondas Curtas (do inglês *visible to near-infrared and short-wave infrared*)
VNIR	Comprimentos de onda do Visível ao Infravermelho Próximo (do inglês *visible to near-infrared*)
V-SWIR	Comprimentos de onda do Visível ao Infravermelho de Ondas Curtas (do inglês *visible to near-infrared and short-wave infrared*)

Estatística

K-folds	Modelo de validação de dados
N_s	*Number of Samples*
OOB	*Out-Of-Bag*
R^2	*Coefficient of Determination*
RMSE	*Root Mean Squared Error*
RPD	*Ratio of Performance to Deviation*
RMSECV	*Root Mean Square Error of Cross-Validation*
RMSEP	*Root Mean Square Error of Prediction*

Instituições

ASD	Fabricante do espectrorradiômetro FieldSpec

CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
ESA	Agência Espacial Europeia (do inglês *European Spatial Agency*)
FAPUR	Fundação de Apoio à Pesquisa Científica e Tecnológica da UFRRJ
IBGE	Instituto Brasileiro de Geografia e Estatística
INPE	Instituto Nacional de Pesquisas Espaciais
SiBCS	Sistema Brasileiro de Classificação de Solos ou *Brazilian System of Soil Classification*
UFRRJ	Universidade Federal Rural do Rio de Janeiro

Localidade

INP	Parque Nacional de Itatiaia (do inglês *Itatiaia National Park*)
RJ	Rio de Janeiro

Pré-processamento espectral

AB-log	*Conversion to absorbance -log10(R)*
CR	*Continuum Removal*
IRF4	Pré-processamento espectral: Inverso da Reflectância por um Fator de 10^4 ou *Inverse of Reflectance by a Factor of* 10^4
SVG	Savitzky-Golay
SVG-1-2-9	*Savitzky–Golay 1st derivative using a 2nd-order polynomial and search window 9*
SVG-1-2-11	*Savitzky–Golay 1st derivative using a 2nd-order polynomial and search window 11*
SVG-1-2-11 + IRF4	*Savitzky–Golay 1st derivative using a 2nd-order polynomial and search window 11 + the Inverse of Reflectance by a factor of* 10^4
IRF4 + SVG-1-2-11	*Inverse of Reflectance by a Factor of* 10^4 *+ Savitzky–Golay 1st derivative using a 2nd-order polynomial and search window 11*
SVG-1-2-11 + IRF4 + NR 434	*Savitzky–Golay 1st derivative using a 2nd-order polynomial and search window 11 + Inverse of Reflectance by a Factor of* 10^4 *+ Noise Reduction* (from 434 nm)
IRF4 + SVG-1-2-11 + NR 434	*Inverse of Reflectance by a Factor of* 10^4 *+ Savitzky–Golay 1st derivative using a 2nd-order polynomial and search window 11 + Noise Reduction (from 434 nm)*
SVG-1-2-11 + NR 434	

	Savitzky–Golay 1st derivative using a 2nd-order polynomial and search window 11 + Noise Reduction (from 434 nm)
IRF4 + NR 434	*Inverse of Reflectance by a Factor of* 10^4 *+ Noise Reduction (from 434 nm)*
PCAL	*Principal Component Analysis Location*
RHCC	*Removal of High Correlated Covariates*
stepAIC	*Stepwise Algorithm Akaike information criteria*

Química e elementos químicos

Al	Alumínio
C	Carbono
Ca	Cálcio
K	Potássio
Mg	Magnésio
Na	Sódio
P	Fósforo
pH	Potencial hidrogeniônico
H	Hidrogênio
N	Nitrogênio
TC	Carbono Total no Solo (do inglês *Total Carbon*)

Sensores

3A60_41	Imagem do sensor CHRIS
3A61_41	Imagem do sensor CHRIS
4CDF_41	Imagem do sensor CHRIS
CHRIS	Satélite Imageador Compacto de Alta Resolução (do inglês *Compact High Resolution Imager*)
PROBA	Satélite que transporta o sensor CHRIS (do inglês *space platform Project for On Board Autonomy*)

Mapeamento e Sensoriamento remoto

DSM	Mapeamento Digital de Solos (do inglês *Digital Soil Mapping*)
HSM	Mapeamento Hiperespectral de Solos (do inglês, *Hyperspectral Soil Mapping*)
RS	Sensoriamento remoto (do inglês *Remote Sensing*)

SUMÁRIO

1 GENERAL INTRODUCTION

Soil is beyond a natural resource which provides energy, food, fiber and many environmental services like water filtering and regulation. Soil supports all terrestrial life. There is an association between soil quality and human health; this association is named soil health, where soil organic carbon is an indicator of soil quality (health). The current human deficiency of essential micro and macro-nutrients must be supplied through soil (LAL, 2016). These reasons only already surplus the agronomic ones in terms of the importance of soils.

Traditionally, the identification and quantification of soil elements (chemistry analyses) are effectuated through titration (WALKLEY; BLACK, 1934; BECKWITH, 1959; DONAGEMA *et al.*, 2011; TEIXEIRA *et al.*, 2017). These methods uses chemical reagents, which must be correctly disposable, increasing their cost and potential of pollution if incorrectly disposed (PAIM *et al.*, 2002; IMBROISI *et al.*, 2006). Alternative methods, such as the application of Remote Sensors, were developed and currently enhanced to bring accurate results. In the context of Proximal Soil Sensing (PSS), Visible-Near-Infrared and Middle Infrared ranges, are the most common (ROSSEL *et al.*, 2009; ROSSEL *et al.*, 2011; NAWAR *et al.*, 2017).

In the perspective of land management of soil properties, the spatial analyses such as mapping is one the most useful tools (LIMA *et al.*, 2013). The Digital Soil Mapping (DSM) with computational and Remote Sensing (RS) resources from satellites has been largely applied to map soil classes and soil properties (LIMA *et al.*, 2013; LAMICHHANE *et al.*, 2019).

Environmental reference areas, natural or without human interference for a long time, are key to test and reply RS and DSM methods. The Itatiaia National Park (INP) is located in the southeastern region of Brazil in the south of Rio de Janeiro state and it has boundaries with Minas Gerais state. It is a conservation unit, the first national park created in 1937, and has a low degree of anthropic actions due to recent agricultural activities. In the upper part of INP, due to the altitude and climatic factors, the soils tend to preserve more carbon than in the highly weathered soils of the lower part of the park. The stratification caused by climatic variation with altitude results in plenty of endemic and threatened of extinction species of fauna and flora (about 116 birds, 73 mammals, and 40 vegetation endemics species). Thus, INP is key to preserve part of Brazilian and Atlantic forest biodiversity (BARRETO *et al.*, 2013b; BARRETO *et al.*, 2013a). The importance of the park is also given by the potential for water distribution in 12 important regional hydrographic basins. Especially in summer, the soils in the INP plateau store rainwater (AXIMOFF *et al.*, 2014), which is slowly released over a long period throughout the year, thus feeding several springs that will contribute to important rivers like the Paraíba do Sul and Rio Grande, even during the dry season.

This work consists of two chapters that are intertwined in the following form: Chapter I - The application of PSS in the upper part of Itatiaia National Park, aiming the spectral record of these soils with the respective horizons, the prediction of their properties using such reflectance spectra, and the exploration and development of spectral preprocessing techniques. Chapter II - Mapping of total carbon in the soil with the technique of combining Multi-Hyperspectral orbital remote sensing images and PSS. The chapters were written in English as they will be submitted for publication in international journals of RS and DSM.

This work has a multidisciplinary approach and includes the areas: i) soil science; ii) data science; iii) mathematical modeling through machine learning algorithms; iv) orbital and proximal remote sensing; and v) DSM.

2 CAPÍTULO I

STUDY OF VIS-NIR SPECTRAL PREPROCESSING AND MACHINE LEARNING ALGORITHMS FOR PREDICTION OF CHEMICAL SOIL PROPERTIES

2.1 RESUMO

A avaliação da refletância na faixa do visível, infravermelho próximo e infravermelho de ondas curtas (Vis-NIR-SWIR, ou simplesmente V-SWIR) combinada com algoritmos constitui métodos aplicáveis para análises do solo. Os objetivos deste estudo foram investigar a capacidade desses métodos de predição de propriedades do solo, tais como o conteúdo dos elementos extraíveis Al, Ca, K, Mg, Na e P; valores de pH, carbono total (TC), H e N; e desenvolver, combinar e comparar diversos pré-processamentos espectrais sobre os comprimentos de onda (350–2500 nm), para serem utilizados nos seguintes algoritmos de aprendizado de máquina: Redes Neurais Artificiais (ANN, do inglês *Artificial Neural Network*), *Random Forest* (RF), *Cubist* (CB), e a comumente aplicada, Regressão de Mínimos Quadrados Parciais (PLSR, do inglês *Partial Least Squares Regression*). Foram coletadas 300 amostras de horizontes de 84 perfis de solo da parte alta do Parque Nacional de Itatiaia (INP, do inglês *Itatiaia National Park*), que está localizado no Estado do Rio de Janeiro, região sudeste do Brasil. O INP é uma unidade de conservação, sendo sua principal finalidade a preservação de fauna e flora, com a possibilidade de atividades de lazer em espaços definidos; portanto, é uma área de referência para estudos do ambiente de Floresta Atlântica. A validação cruzada do tipo k-*fold* foi implementada para dividir os dados e validar 6.000 modelos. O melhor pré-processamento espectral foi o Inverso de Refletância de Fator de 10^4 (IRF4) para TC com algoritmo CB, superando os métodos de pré-processamentos comumente utilizados, com um R^2 médio entre os *folds* de 0,85, RMSE de 1,96; e 0,67 com 0,041 respectivamente para H. Dentro dos *folds* preditos com os modelos para TC, o melhor teve valor R^2 de 0,95. A boa correlação com as técnicas V-SWIR mostra que o método pode ser usado para predição de propriedades de solos e, do mesmo modo, utilizado para o monitoramento ambiental rápido.

Palavras-chave: Pedometria. Modelagem de solo. Sensoriamento remoto proximal.

2.2 ABSTRACT

Evaluation of visible, near-infrared and shortwave infrared reflectance (Vis-NIR-SWIR, for short V-SWIR) combined with algorithms constitute applicable methods for soil analysis. The main objectives of this study were to investigate the ability of these methods to predict soil properties, such as the contents of the extractable elements, Al, Ca, K, Mg, Na, P, and values of pH, total carbon (TC), H and N. Also, to develop, combine and compare different spectral (350–2500 nm) preprocessing to be applied in machine learning algorithms such as: Artificial Neural Network (ANN), Random Forest (RF), Cubist (CB) and the common Partial Least Squares Regression (PLSR). A total of 300 soil samples from horizons of 84 soil profiles were collected in the upper part of Itatiaia National Park (INP), located in the State of Rio de Janeiro, Southeastern Brazil. The INP is a conservation unit, and its main mission is the preservation of fauna and flora, with some areas set for leisure; thus it is a reference area for environmental studies of the Atlantic Forest. The k-fold cross validation approach was implemented to split data and validate 6,000 models. The best spectral preprocessing was the Inverse of Reflectance to Factor of 10^4 (IRF4) for TC with CB algorithm, outperforming the commonly used preprocessing methods, with an averaged R^2 among the folds of 0.85, RMSE of 1.96; and 0.67 with 0.041 respectively for H. Within the predicted folds with the models of TC, the highest prediction had a R^2 of 0.95. The good correlation with V-SWIR techniques shows that the method can be used for soil properties prediction, and used for fast environmental monitoring.

Keywords: Pedometrics. Soil modeling. Proximal soil sensing.

2.3 INTRODUCTION

Soils in tropical regions are predominantly highly weathered and they usually have low organic carbon (C) and nitrogen (N) in the upper horizons. However, in the high altitudes of mountain ranges, peculiar climate with low temperatures and endemic vegetation occur, resulting in distinct soil formation processes. The Itatiaia National Park (INP) in the State of Rio de Janeiro, Southeastern region of Brazil, is an example of these conditions, mainly in the INP upper part (plateau), which is above 2000 msnm of the topographic contour line. The climatic conditions and rock outcrops favor the occurrence of herbaceous graminoid plants, mostly *Cyperaceae* and *Poaceae*, arranged in clumps, with few incidences of other species (SOARES *et al.*, 2016). The low temperatures also lead to preservation of C and its incorporation into the soil matrix. Thus, compared to most tropical soils, the INP soils have a large amount of Total Carbon (TC), which reaches up to 29.5%, according to recent studies of Costa *et al.* (2020) in the INP, and thus an elevated content of N, due to the strong correlation between them. The mountainous relief and the access limited to field trails in the INP plateau, results in hard locomotion, transport and poor communications, all making it more difficult for field campaigns to map soils and other studies as well.

Under the umbrella of the Remote Sensing tools, the Proximal Soil Sensing (PSS) has been developed and applied to predict soil properties features. Different processing algorithms and methods, such as preprocessing spectral algorithms, statistical predictions and machine learning algorithms, are able to deliver different accuracy for each soil property and feature. With appropriate management the prediction of soil organic carbon using spectral data can reach high accuracy (DANGAL *et al.*, 2019). The main contribution from this study is to bring a wide range of combinations of spectral preprocessing techniques with statistical and machine learning algorithms for different soil properties prediction using spectral reflectance.

This study was motivated by the fact that the INP plateau has peculiar and distinct environmental conditions from most tropical regions of Brazil, resulting in soils with high organic carbon contents, consequently, different soil types, properties and environmental services. The fact that the INP is a conservation unity, makes it a reference study area for Atlantic Forest ecosystems, such as the altitude fields that predominate in the plateau region. In addition, the registration of spectrum of soils from the INP plateau will allow for further studies, as the methods and techniques of V-SWIR spectrum processes evolves, and to monitor variations in the soil properties as a result of changes in climate behavior.

The hypothesis of the work is that the combination of spectral preprocessing techniques can enhance the prediction of soil properties across different statistical and machine learning methods using spectral data. Also, that the preprocessing application always improve the prediction capacity.

The general objective of the study was to test and compare the capacity of ANN, PLSR, RF and CB algorithms for predicting the contents of the extractable elements aluminum (Al), calcium (Ca), potassium (K), magnesium (Mg), sodium (Na), phosphorus (P); values of pH, total carbon (TC), hydrogen (H) and nitrogen (N); through the development, combination and application of spectral preprocessing techniques along the V-SWIR spectral region.

2.4 LITERATURE REVIEW

2.4.1 The Reflectance

According to Jensen (2014), reflectance is beyond the simple “process whereby radiation “bounces off”an object like [...] bare soil”. Indeed, the process involves “reradiation of photons in unison by atoms or molecules in a layer approximately one-half wavelength deep”. The reflecting surfaces can be from essentially smooth to roughness, respectively acting as a perfect specular reflector, or perfect diffuse reflector, also called Lambertian surface where “the radiant flux leaving the surface is constant for any angle of reflectance”. To illustrate that, a very calm water surface acts like specular reflector, while forest diffuse reflector, and a material like spectralon (material made from polytetrafluoroethylene) is nearly perfectly Lambertian, with the reflectance is generally more than 99% over a range from 400 to 1500 nm and more than 95% from 250 to 2500 nm.

The radiation budget equation (Equation 2.1) says that the total amount of radiant flux in specific wavelengths (λ) incident to the terrain (or surface) (Φ_{i_λ}), from any angle in a hemisphere, must be accounted for by evaluating the amounts of: radiant flux reflected from the surface ($\Phi_{reflected_\lambda}$), radiant flux absorbed by the surface ($\Phi_{absorbed_\lambda}$), and radiant flux transmitted through the surface ($\Phi_{transmitted_\lambda}$):

$$\Phi_{i_\lambda} = \Phi_{reflected_\lambda} + \Phi_{absorbed_\lambda} + \Phi_{transmitted_\lambda} \tag{2.1}$$

As soil is an opaque object (where PSS wavelengths can not pass through), consequently transmittance is zero, the radiation budget equation (Equation 2.2) can be write as:

$$\Phi_{i_\lambda} = \Phi_{reflected_\lambda} + \Phi_{absorbed_\lambda} \tag{2.2}$$

Thus, the Hemispherical reflectance (ρ_λ) (dimensionless) is defined as the “ratio of the radiant flux reflected from a surface to the radiant flux incident to it”, given by Equation 2.3:

$$\rho_\lambda = \frac{\Phi_{reflected_\lambda}}{\Phi_{i_\lambda}} \tag{2.3}$$

In remote sensing the spectrum curves are usually presented percent reflectance. In fact, if we take the simple hemispherical reflectance equation and multiply it by 100, we obtain an expression for percent reflectance ($\rho_{\lambda_\%}$), given by Equation 2.4:

$$\rho_{\lambda_\%} = \frac{\Phi_{reflected_\lambda}}{\Phi_{i_\lambda}} \times 100 \tag{2.4}$$

In practical aspects, the reflectance spectrum is computed by dividing the spectral response of the target of interest (soil, vegetation, any surface or object) by the spectral response of the reference sample (commonly spectralon).

2.4.2 Wavelengths Intervals on Proximal Soil Sensing

The Proximal Sensing techniques have been applied in a wide range of fields (LU; FEI, 2014), including soil science. The Proximal Soil Sensing (PSS) techniques are fast, non-destructive, environmental-friendly and cost-effective (ROSSEL *et al.*, 2005). PSS deals with wavelength range between 350 and 30,000 nm, visible and infra-red regions respectively. There is no consensus about wavelength segments nomenclature (FANG *et al.*, 2018). According to Jensen (2014) the wavelength interval in the electromagnetic spectrum are commonly referred to as a *band*, *channel*, or *region*; delimited by the channels: Violet limit 400 nm; Blue 450 nm; Green 500 nm; Green limit 550 nm; Yellow 580 nm; Orange 600 nm; Red 650 nm; Red limit 700 nm; Near-infrared 1000 nm; Far-infrared 30000 nm. By intervals as: Ultra Violet (UV, 254 to 400 nm); Visible (Vis, 400 to 700 nm); Near Infrared (NIR, 700 to 1.000 nm) and Middle-Infrared region (MIR, often referred to as the short wavelength infrared, SWIR) includes energy with a wavelength of 1300 to 3000 nm.

The branch of knowledge of PSS studies presented the spectral range as: VNIR spectrum (400-1100 nm) (DANIEL *et al.*, 2003); VIS (400-700 nm), NIR (700-2500 nm) and MIR (2500-25,000 nm) (ROSSEL *et al.*, 2005); VIS (400-700 nm), NIR (700-1100 nm), and SWIR (1100-2500 nm) (BEN-DOR *et al.*, 2009); Vis-NIR (350-2500 nm) and MIR (2500-25,000 nm or 4000-400 cm^{-1}, as Wave Number expressed in terms of energy) (TERRA *et al.*, 2015). Where Wave Number (Ψ) is the number of waves in a unit length (usually per cm) (JENSEN, 2014); Vis (350-780 nm); NIR (780-2.500 nm); VNIR (350-1000 nm) and Vis-NIR (350-2.500 nm)) (FANG *et al.*, 2018); NIR (700-2.500 nm) and MIR (2.500-25.000 nm) (DANGAL *et al.*, 2019); Vis-NIR (400-2500 nm) (ROSSEL *et al.*, 2009; GOMEZ *et al.*, 2012; VASQUES *et al.*, 2014; NOURI *et al.*, 2017; PADARIAN *et al.*, 2019); Vis-NIR (350-2500 nm) (CONFORTI *et al.*, 2015; MCGILL *et al.*, 2015; TERRA *et al.*, 2015; DOTTO *et al.*, 2018; PINHEIRO *et al.*, 2017; ROUDIER *et al.*, 2017); Vis-SWIR (350-2.500 nm) (CHICATI *et al.*, 2019); VIS-NIR-SWIR (400-2500 nm) (DEMATTÊ *et al.*, 2016).

In agreement with Jensen (2014), Demattê *et al.* (2016), Chicati *et al.* (2019), in this study we adopted the definition of Visible, Near-Infrared and Shortwave Infrared reflectance (Vis-NIR-SWIR, for short V-SWIR) as ranging from 350 to 2500 nm.

2.4.3 Proximal Sensing to Proximal Soil Sensing and its Features

One of the early spectral studies started around 1900s with optical properties of Iodine (COBLENTZ, 1903) and Infra-Red absorption spectra of organic compounds (COBLENTZ, 1904). Later it was applied to inorganic and mineralogical fields (HUNT *et al.*, 1950; HUNT; TURNER, 1953). A chart with probable positions of characteristic Infra-Red absorption bands of organic and inorganic compounds was presented by Colthup (1950).

Among the initial spectral soil investigations we find the studies of Bowers & Hanks (1964), who analyzed the soil reflectance according to moisture content, organic matter, and particle size. The vibrational – electronic processes with the spectral signatures of soil minerals was early characterized by Hunt (1977), with some recent studies by Fang *et al.* (2018) and Chicati *et al.* (2019). Soil properties prediction using V-SWIR and MIR has been a topic of development for the PSS field. Beyond V-SWIR, MIR has shown a better capacity to predict soil properties (CLAIROTTE *et al.*, 2016; DANGAL *et al.*, 2019).

According to Bowers & Hanks (1964) “The elevated daytime temperatures of dark-colored soils is attributed to their greater absorption of solar radiant energy”. Soil organic matter is the main soil constituent that contribute to the dark soil color. Organic matter (OM) and iron oxides tend to absorb incident radiation relatively homogeneously across V-SWIR wavelengths and a low iron content changes the shape of the curve from a horizontal to a positive upward trend (DEMATTÊ, 2002), this is confirmed when the OM is removed (DAS *et al.*, 2015); and the absorption power of incident radiation of OM is mainly between 450 and 1000 nm (FORMAGGIO *et al.*, 1996).

Spectral reflectance increases with the decrease in the size of soil particles (BOWERS; HANKS, 1964), especially below 0.4 mm in diameter (DAS *et al.*, 2015). Correlations among soil attributes and spectral signatures were found by Demattê *et al.* (2012), Chicati *et al.* (2019), and these spectral characteristics allow the PSS to identify the constituents of the soil (CHICATI *et al.*, 2019). Spectral behavior of different soil classes is governed mainly by mineralogical composition, OM content, and grain size (FORMAGGIO *et al.*, 1996). In a study of spectral behavior of a Brazilian soil in which the OM was removed by chemical treatment, there was an increase of the reflectance factor by more than 100%, driven mostly by the interaction of OM with oxide minerals (GALVÃO; VITORELLO, 1998).

Among the spectral reflectance most studied soil properties we find contents of Al, H, Ca, K, Mg, Na, N, P, pH, OM; iron oxides, and granulometry (FORMAGGIO *et al.*, 1996; ROSSEL *et al.*, 2005; GOMEZ *et al.*, 2008; DEMATTÊ *et al.*, 2012).

Another approach is based on the fact that soil has its own spectral behavior also called as spectral signature, which varies according to its mineralogical composition, the content of organic matter, moisture and granulometry (DAS *et al.*, 2015). For the spectral comparison techniques, it is useful to create spectral libraries. Many institutions have their own spectral libraries (MULDER *et al.*, 2011); such as the Brazilian Spectral Library, which can contribute to the prediction of soil properties such as clay, sand, organic matter, cation exchange capacity, pH and base saturation in soil (DEMATTÊ *et al.*, 2019). They consist of reference spectra, obtained in laboratories or in the field under more controlled conditions (GALVÃO *et al.*, 2001; MULDER *et al.*, 2011; DEMATTÊ *et al.*, 2019). The importance of spectral libraries was also highlighted in a study with spectral data from soils from several European countries (PADARIAN *et al.*, 2019).

Although it is possible to identify soil constituents through spectral behavior, in other words analyzing the curves and their features like intensity and shapes (DEMATTÊ *et al.*, 2012), the spectral data is generally large data, being hard for the human mind to compile. The application of computational algorithms and software (ROSSEL, 2008) can heighten this capacity to predict information from spectral data. Several soil properties and attributes have been predicted using different methods and predictive algorithms (ROSSEL *et al.*, 2005).

2.4.4 Algorithms and Preprocessing

The most applied technique to predict soil properties is the Partial Least Squares Regression (PLSR), which is a statistical method (GOMEZ *et al.*, 2008; ROSSEL *et al.*, 2008; ADELINE *et al.*, 2017; KOPAČKOVÁ *et al.*, 2017; NANNI *et al.*, 2018). Machine Learning Algorithms such as Artificial Neural Network (ANN) were also applied with satisfactory results (DANIEL *et al.*, 2003; ROSSEL; BEHRENS, 2010). Other studies compared the ability to predict soil properties by spectral data using algorithms such as ANN, Random Forest (RF), PLSR and Cubist (CB) (MOUAZEN *et al.*, 2010; KUANG *et al.*, 2015; MORELLOS *et al.*, 2016; NAWAR *et al.*, 2017; DANGAL *et al.*, 2019).

Although raw spectral data may be used for soil property prediction (CHICATI *et al.*, 2019), the use of preprocessing can enhance the algorithms' capacity. Some of common used techniques as the Savitzky-Golay filter (SAVITZKY; GOLAY, 1964) to the noise reduction, and studies showed that it improves the results of soil properties prediction (MALEKI *et al.*, 2006; CLAIROTTE *et al.*, 2016; GHOLIZADEH *et al.*, 2016; CEZAR *et al.*, 2019). The Continuum Removal approach has also been used with good results (CLARK, 1999; TERRA *et al.*, 2015; KOPAČKOVÁ *et al.*, 2017; PINHEIRO *et al.*, 2017).

2.5 MATERIAL AND METHODS

2.5.1 Description of the Study Area

The INP was the first national park created in Brazil (1937). It is designated as an area for nature conservancy and preservation of the Atlantic Forest biome, due to its high biodiversity. Since the eighteenth-century researchers from many countries have visited the region to study the biodiversity of the *Serra da Mantiqueira* mountains, where the INP is located.

The park has an area of 28035 hectares (blue line in Figure 2.1), which is divided into three sectors, Mauá (east), the lower part (south) and the upper part with 16402 hectares (central and northwest, green line in Figure 2.1). The study was conducted in the upper part of the park, region designated by Tomzhinski *et al.* (2012) to those areas above the 2000 m of the topographic line to the south, southeast and southwest, going toward the limits of the INP in the other directions. The highest point in the park is the so-called *Pico das Agulhas Negras* with an elevation of 2,791.6 msnm. The INP is located at the border of Rio de Janeiro and Minas Gerais states, and is also near the São Paulo state border in Southeast region of Brazil. The area is comprised by the UTM coordinates 523500–546500m E and 7514500–7540500m N, Zone 23 K, datum WGS84.

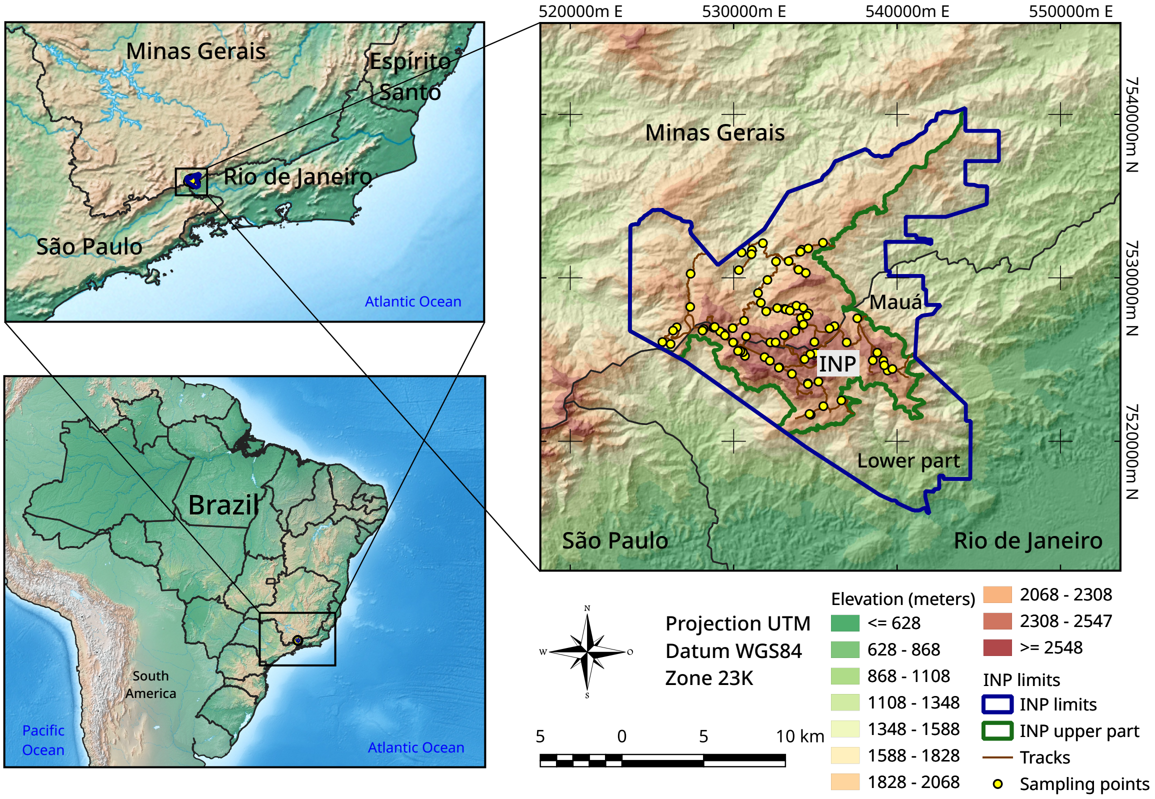


Figure 2.1: Itatiaia National Park and upper part (plateau) location, near the triple border of Rio de Janeiro, Minas Gerais and São Paulo states. To the right, the three sections of the park: Mauá (east), lower part (south) and the upper part (central and northwest). Source of area delimitation lines: IBGE (2010), INP managers including Tomzhinski *et al.* (2012).

The vegetation profile inside the park changes with the altitude, composed of Dense Ombrophilous Forest, which is split in three sub categories: Sub-Montane: located in plateau slopes until 500 m; Montane: in the higher part of the plateau from 500 m to 1.500 m of altitude; and

High-Montane: above limits of Montane (1.500 m). The Montane Mixed Ombrophilous Forest is composed of vegetation that has the upper extract is mostly populated by the conifer *Araucaria angustifólia* a pioneer specie. The main characteristic of the Montana Semi Deciduous Seasonal Forest is presented by the fall of the leaves between 20 and 50% of the total, located from 400 to 1500 m of altitude. Transition systems: when mutual incursion of flora verified. And Vegetation Refuge (17,08%): in the highest altitudes, usually above 1500 m, are the Altitude Fields (BARRETO *et al.*, 2013a), mostly composed of herbaceous graminoid plants (SOARES *et al.*, 2016). In proportions: High-Montane Ombrophilous Forest 50.17%; Vegetation Refuge 17.08%; Montane Ombrophilous Forest 12.09%; Rock outcrops 3.55%; Other 17.11%; (BARRETO *et al.*, 2013a).

According to the Köppen classification, the climate in the INP is Cwb-type (MODENESI, 1992; SANTOS *et al.*, 2000; ALVARES *et al.*, 2013). The annual average temperature is 11.5°C, and the average during winter is 8.4°C, sporadically reaching below zero. The dominant geology of the INP plateau is formed by alkaline syenites and granite-gneissic rocks (BARRETO *et al.*, 2013a; ROSA; RUBERTI, 2018). The landscape is mainly formed by high mountains and escarpments with narrow valleys among the rock outcrops. The INP plateau has an expressive area of soils with an organic horizon, covered mostly by herbaceous graminoid plants, with a predominance of *Cyperaceae* and *Poaceae* arranged in clumps (SOARES *et al.*, 2016), located in the narrow valleys and talus with lower slopes.

2.5.2 Soil Sampling, Analysis and Preparations

Since the access to the area is restricted, due to the presence of endemic species, environmental protection requirements, steep terrain and short number trails, the Conditioned Latin Hypercube Sampling (cHLS) algorithm (MINASNY; MCBRATNEY, 2006) was set to place soil sampling locations near the trails, with a 100 m buffer from the paths with highest potential to express the variability of the soils in the region, as applied by Costa *et al.* (2020).

Initially, 80 sampling points were determined, but 6 fell on rock outcrops. During the field excursion, 10 random sample points were added aiming to cover the range of INP soil variability, based on the experience of the research team, resulting in a total of 84 profiles (presented in Figure 2.1 as yellow dots).

Among those 84 profiles, as describe by Costa *et al.* (2020), 33 were classified as *Organossolos* according to the Brazilian System of Soil Classification (SiBCS) (SANTOS *et al.*, 2018), which is an equivalent of Histosols (IUSS Working Group WRB, 2015); 34 *Cambissolos* (subdivided as 25 *CAMBISSOLO HÚMICO* and 9 *CAMBISSOLO HÍSTICO*) equivalent of Cambisols, although both classes correspond to a top layer rich in OM; and 13 *NEOSSOLOS* equivalent of Leptosols (in this case), with 5 of these shallow soils also having surface horizons rich in OM. The last four profiles (one Alisol, one Ferralsol and two Cambisols) have a lower content of OM.

From the horizons of the 84 profiles, 300 soil samples were obtained. Among the 300 soil samples, 96 were from horizons with high carbon content. The samples were prepared to obtain the fine earth fraction and after that chemically analyzed for Al, H, Ca, K, Mg, Na, N, P, pH and TC, according to Teixeira *et al.* (2017), and also used to analyze with spectral readings, as described below.

Aiming to neutralize the influence of moisture on the spectral reading, soil samples were placed in paper bags then dried in an oven under forced air circulation, at a temperature of 45°C for 48 hours (DEMATTÊ *et al.*, 2012; DEMATTÊ; TERRA, 2013; TERRA *et al.*, 2018). When removed from the oven the samples were placed in a glass desiccator until they reached ambient temperature (close to 25 Celsius). Figure 2.2 presents a summary of these steps.

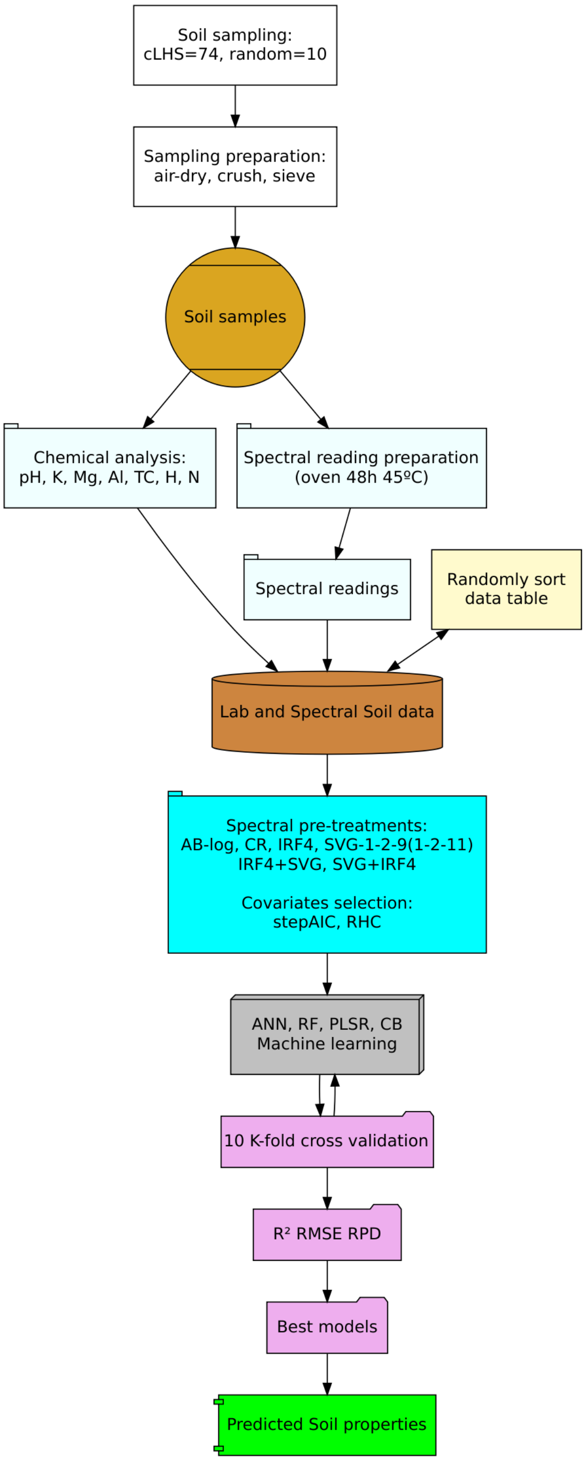
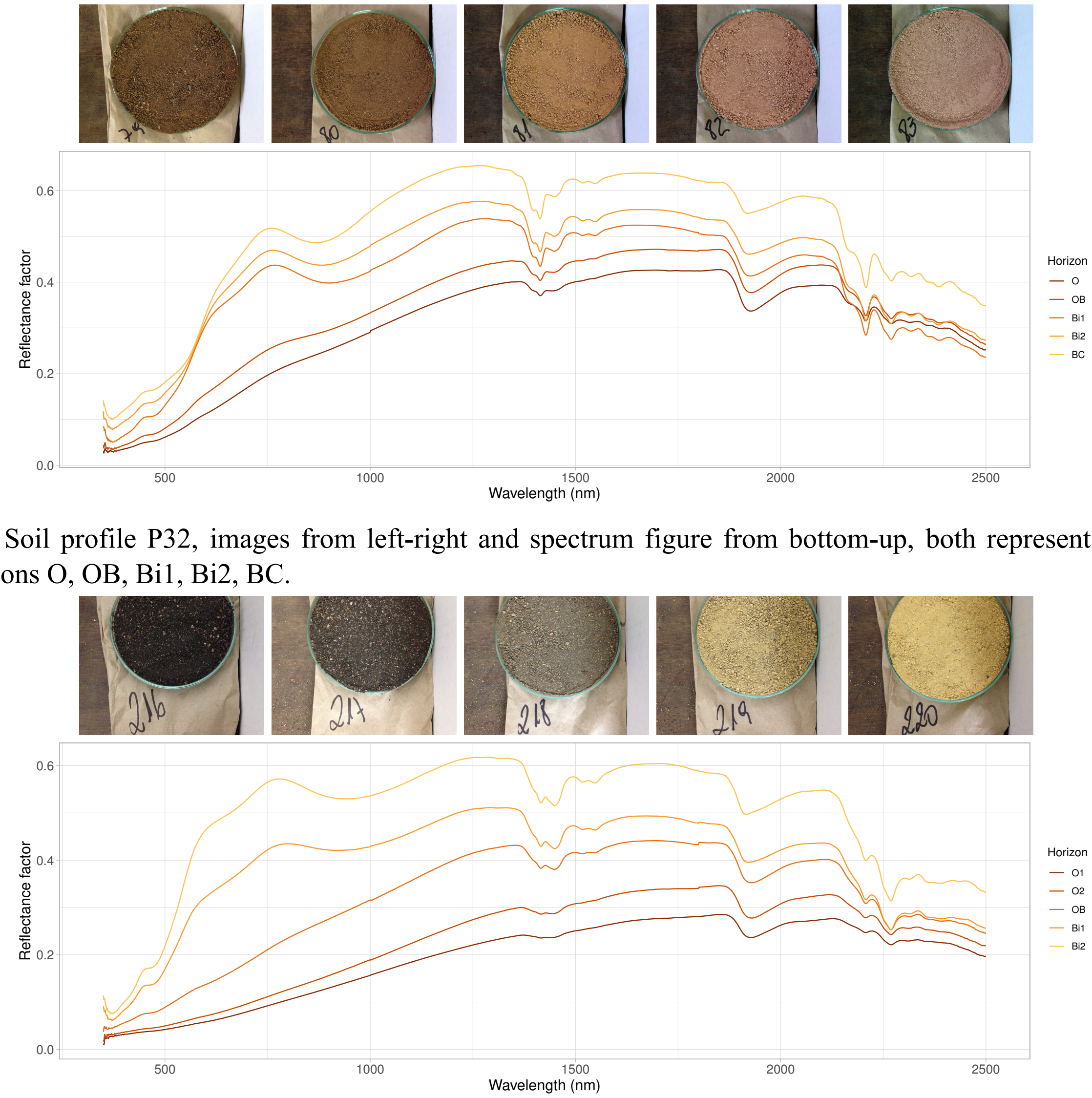


Figure 2.2: Workflow of the soil analysis and spectral predictions of soil properties.

For the spectral readings, soil samples were placed into Petry dishes of 9 cm in diameter, just before the reading, to avoid absorption of ambient moisture. The soil samples were directly placed on the dishes, without mixing, causing a heterogeneous particle distribution, where the finer particles stayed on the surface, generally distributing from the center. Following the preparation chain, the sample surface was gently compacted with a flat circular glass object, to obtain a leveled and flat surface (Figure 2.3). They were then read with an ASD FieldSpec 4 spectrometer with the following characteristics: Spectral Range from 350 to 2500 nm (V-SWIR); Spectral Resolution of 3 nm @ 700 nm and 10 nm @ 1400/2100 nm; Spectral sampling (bandwidth) 1.4 nm @ 350-1000 nm and 1.1 nm @ 1001-2500 nm. All spectral readings were conducted in a dark room and on the same day. To avoid light source oscillations and consequent variations between readings, a battery powered no-break line was connected to the devices.

The light source was a 70-Watt halogen bulb lamp, positioned 15° from nadir, at a distance of 70 cm. The optical fiber probe sensor was placed 35 cm from the soil samples with an objective lens of 8° (lens angle), positioned at nadir (0°) in relation to the soil samples. Each soil sample underwent 100 scans. During the scans, three 120° rotations were executed to attain a homogeneous reading over the entire surface of each sample. At every 30 minutes, or 24 samples, the optimize and white reference were read according to equipment manufacturer instructions.



(a) Soil profile P32, images from left-right and spectrum figure from bottom-up, both represents the horizons O, OB, Bi1, Bi2, BC.

(b) Soil profile P77, images from left-right and spectrum figure from bottom-up, both represents the horizons O1, O2, OB, Bi1, Bi2.

Figure 2.3: Reflectance of profiles P32(a) and P77(b) horizons, both classified as *CAMBISSOLO HÍSTICO Distrófico típico* (SiBCS), equivalent to Histic Cambisols (WRBSRG). Images show the flattening sample surface.

2.5.3 Data Handling, Spectral Preprocessing, Covariates Selection

The spectroradiometer data (format file .asd) was converted in a plain text file (with 15 decimals after the comma), then, as data-table, associated to the contents of Al, Ca, H, K, Mg, Na, N, P, and values of pH and TC data, from the wet chemistry laboratory analyses. To improve the prediction results, two approaches were adopted. The first approach used in the spectral data such as Continuum Removal (CR) (CLARK, 1999), Savitzky-Golay (SVG) (SAVITZKY; GOLAY, 1964) with different settings across the derivative, order polynomial and search window (VASQUES *et al.*, 2008), and Inverse of Reflectance to Factor of 10^4 (IRF4). The IRF4 was obtained dividing 10,000 for each value of the reflectance spectrum. A conversion of spectral data to absorbance by the -log10 (reflectance) (ROSSEL *et al.*, 2005) (AB-log) was also included as a preprocessing (Table 2.1) (Figure 2.4). The second approach used techniques for dimensionality reduction of spectral covariates, such as the Stepwise Algorithm by the Akaike information criteria (stepAIC), which removed 1851 of the 2150 covariates, keeping ~14% (299 covariates). The second technique was the Removal of High Correlated Covariates (RHCC) by the correlation matrix approach, which removed 480 covariates from the dataset keeping 1686.

Table 2.1: Preprocessing applied to spectral data from soil samples of the Itatiaia National Park.

Preprocessing	Abbreviation
Conversion to absorbance -log10(R)	AB-log
Continuum Removal	CR
Inverse of Reflectance to Factor of 10^4	IRF4
Savitzky–Golay 1st derivative using a 2nd-order polynomial and search window 9	SVG-1-2-9
Savitzky–Golay 1st derivative using a 2nd-order polynomial and search window 11	SVG-1-2-11
Savitzky–Golay 1st derivative using a 2nd-order polynomial and search window 11 + the Inverse of Reflectance to factor of 10^4	SVG-1-2-11 + IRF4
Inverse of Reflectance to Factor of 10^4 + Savitzky–Golay 1st derivative using a 2nd-order polynomial and search window 11	IRF4 + SVG-1-2-11
Savitzky–Golay 1st derivative using a 2nd-order polynomial and search window 11 + Inverse of Reflectance to Factor of 10^4 + Noise Reduction (from 434 nm)	SVG-1-2-11 + IRF4 + NR 434
Inverse of Reflectance to Factor of 10^4 + Savitzky–Golay 1st derivative using a 2nd-order polynomial and search window 11 + Noise Reduction (from 434 nm)	IRF4 + SVG-1-2-11 + NR 434
Savitzky–Golay 1st derivative using a 2nd-order polynomial and search window 11 + Noise Reduction (from 434 nm)	SVG-1-2-11 + NR 434
Inverse of Reflectance to Factor of 10^4 + Noise Reduction (from 434 nm)	IRF4 + NR 434

Seeking errors in the dataset (the 300 samples), we also tested the removal of outliers with a Principal Component Analysis Location (PCAL), by taken 10 samples located outside the standard deviation distance of five percent. A similar idea was used by Dangal *et al.* (2019), which evaluated the model performance before and after removing the outlier samples.

The best results from the two approaches were combined then reapplied in the algorithms. For example, after applying SVG we performed IRF4 to the result of SVG and vice-versa, resulting in two different preprocessing SVG-1-2-11+IRF4 (SVG first, with IRF4) and IRF4+SVG-1-2-11 (IRF4 first, with SVG) as displayed in Table 2.1 and Figure 2.4. Noise was also identified and was removed with Noise Removal (350-433 nm) (NR), after being visually identified in the spectral graphs as the initial (83) wavelengths of the IRF4 curve (Figure 2.4).

The dataset (wet chemistry laboratory and spectral combined) was randomly sorted once, to avoid bias due closely samples from the same profile. The K-folds cross-validation method

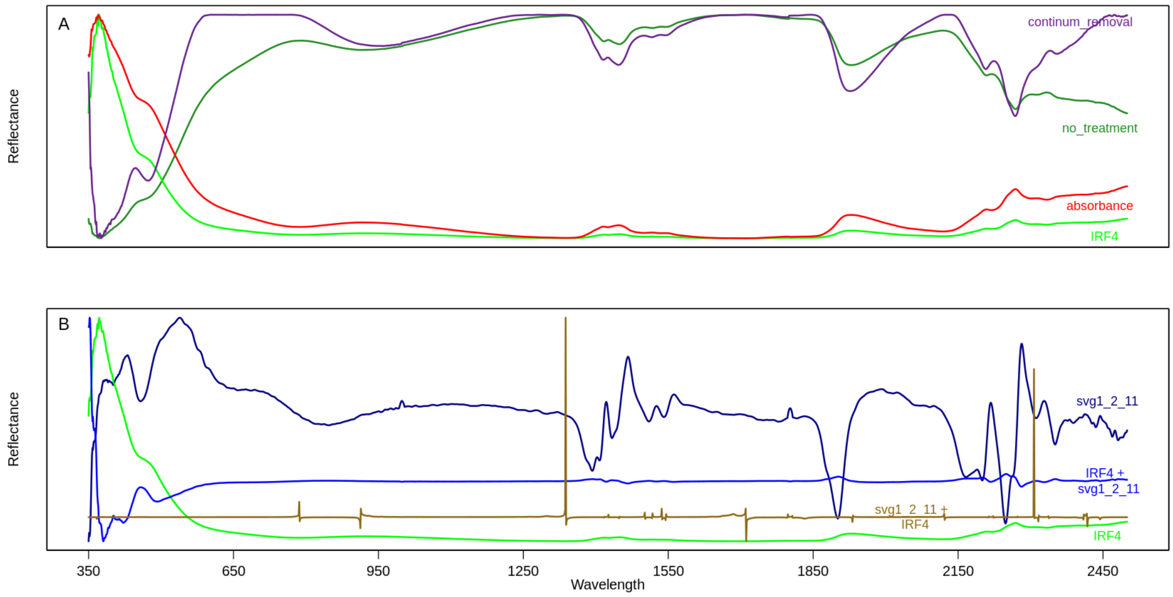


Figure 2.4: Visualization of main Spectral preprocessing curves: (A) Continuum Removal (magenta); no treatment (raw spectrum) (green); absorbance (red); Inverse of Reflectance to Factor of 10^4 (light green). (B) Savitzky-Golay first derivative (dark blue); Inverse of Reflectance to Factor of 10^4 + Savitzky-Golay first derivative (light blue); Savitzky-Golay first derivative + Inverse of Reflectance to Factor of 10^4 (brown); Inverse of Reflectance to Factor of 10^4 (light green). Notice, each curve fits its own y (reflectance) scale.

was implemented. The dataset was submitted to each preprocessing and technique, as defined in Table 2.1. The data from wet chemistry remained unchanged, in other words, it was not treated or converted using any sort of method, only the spectral data was managed through the preprocessing. As a reference, the raw data (with no treatment) was also computed across the models.

2.5.4 Artificial Neural Network (ANN)

The Artificial Neural Network (ANN) was initially developed in 1958 (ROSENBLATT, 1958) and was revised in the 1980s and '90s. A literature review (ROSSEL *et al.*, 2005) has shown that ANN was not commonly applied for prediction of soil properties, PLSR is largely used instead. One of the first applications of ANN for prediction of soil properties (DANIEL *et al.*, 2003) used different bandwidths, 10, 20, 50 and 100 nm.

To apply the ANN algorithm in this study each soil property with the correspondent spectrum from the respective soil sample was scaled between zero (0) and one (1). To apply the other algorithms (RF, PLSR and CB) data was not scaled. After that, the predicted data was converted to the original scale to proceed with the validation.

The ANN consists of an input data layer (all available samples), a hidden layer(s) (of which there may be one or more, in the case of the study, five hidden layers), and one output layer. Combinations were tested, in this case, the number of neurons per layer were defined by Fibonacci sequence, given by the Equation 2.5, where we use n between 7 to 2, giving 13, 8, 5, 3, 1 respectively. In other words, each number is found by adding up the two numbers before it. Thus, the hidden layers are a combination of five layers containing 13-8-5-3-1 neurons respectively. In this arrangement, each neuron was linked with all of the neurons in the next layer, but had no linkage with others neurons in the same layer (Figure 2.5).

$$F_n = \frac{\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n}{\sqrt{5}} \tag{2.5}$$

The implemented ANN was the *resilient backpropagation* type, with R package *neuralnet*, through the following R command `ann_model <- neuralnet(formula = form, data = train_data, hidden = c(13, 8, 5, 3, 1), linear.output = TRUE)`, where the `ann_model` is an object that contains the output ANN model (to be used in the prediction step (further)); `neuralnet` is the R function for ANN (from R package *neuralnet*); `formula = form` is an object that contains the working soil property + covariates (in this study, generally, all and each spectral wavelengths, for example *TC+W350+W351+W352+...+W2150+W2151*); `data = train_data` is the data used to train the model, which the model try to learn the amount of used soil property (TC for example) from spectral data; `hidden = c(13, 8, 5, 3, 1)` the five hidden layers and the respective number of neurons, and the output layer; and `linear.output = TRUE` is the model parameter associated with *act.fct* and *logistic* function.

Then, the soil properties predictions were performed with the R command `fitted_ann <- predict(ann_model, validation_data)`, where the `fitted_ann` contains the the values of each predicted property (for example TC); `predict` is a R "is a generic function for predictions from the results of various model fitting functions. The function invokes particular methods which depend on the class of the first argument" (R Core Team, 2019), this function belongs to R stats package (one of the packages supplied in R base packages set). In this case, the function was applied to a Machine Learning *neuralnet* object (argument) class; `ann_model` as described earlier; `validation_data` is the data that contains only spectral data, where the model tries to apply what was learned before to give the desired prediction. Then the result of the prediction has to be compared with the actual values to assess (validate) the modeling.

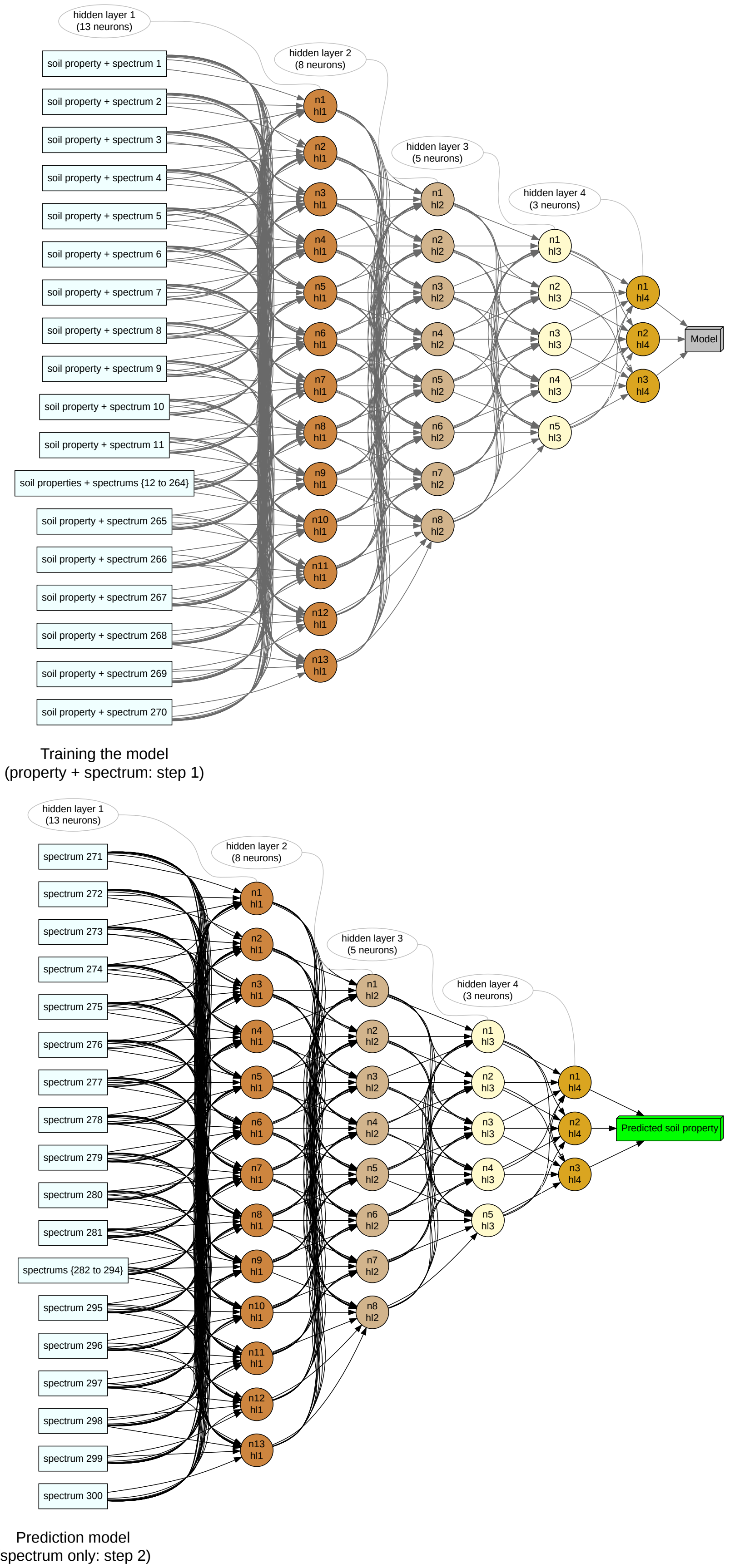
Training and prediction, the algorithm steps, the case for ANN

To train the models, soil property + spectral data were added to the network (Figure 2.5, step 1). This allowed the system to build a weighted connection balanced for the prediction (a model). To perform the prediction, new data (spectral only) from the dataset was injected into the model generating the predicted soil property (Figure 2.5, step 2). The result of the predicted soil property is then placed alongside the observed values to assess the predictive capacity as displayed along the coefficients in Tables 2.3, 2.4, and Figure 2.7 together with Figure 2.8.

The referred above training and prediction processes were repeated for each preprocessing tool, combined to covariate selection methods. Consecutively, these steps were applied for the following algorithms. Detailed information about train and validation aspects are given after the algorithms topics.

2.5.5 Random Forest (RF)

The Random Forest (RF) algorithm (BREIMAN, 2001) is based on regression and classification trees. It builds various regression or classification trees with bootstrap sampling (one third approximately) of the input covariates and internal validation called out of bag (OOB) (LAWRENCE *et al.*, 2006; NAWAR *et al.*, 2017; DANGAL *et al.*, 2019). The model presents the average estimate of the trees for soil attributes prediction (continuous data) and more voted classes for soil types (categorical data). As for ANN and other algorithms, soil plus the covariate data


hidden layer 1 (13 neurons)
hidden layer 2 (8 neurons)
hidden layer 3 (5 neurons)
hidden layer 4 (3 neurons)
soil property + spectrum 1
soil property + spectrum 2
soil property + spectrum 3
soil property + spectrum 4
soil property + spectrum 5
soil property + spectrum 6
soil property + spectrum 7
soil property + spectrum 8
soil property + spectrum 9
soil property + spectrum 10
soil property + spectrum 11
soil properties + spectrums {12 to 264}
soil property + spectrum 265
soil property + spectrum 266
soil property + spectrum 267
soil property + spectrum 268
soil property + spectrum 269
soil property + spectrum 270
n1 hl1
n2 hl1
n3 hl1
n4 hl1
n5 hl1
n6 hl1
n7 hl1
n8 hl1
n9 hl1
n10 hl1
n11 hl1
n12 hl1
n13 hl1
n1 hl2
n2 hl2
n3 hl2
n4 hl2
n5 hl2
n6 hl2
n7 hl2
n8 hl2
n1 hl3
n2 hl3
n3 hl3
n4 hl3
n5 hl3
n1 hl4
n2 hl4
n3 hl4
Model
hidden layer 1 (13 neurons)
hidden layer 2 (8 neurons)
hidden layer 3 (5 neurons)
hidden layer 4 (3 neurons)
spectrum 271
spectrum 272
spectrum 273
spectrum 274
spectrum 275
spectrum 276
spectrum 277
spectrum 278
spectrum 279
spectrum 280
spectrum 281
spectrums {282 to 294}
spectrum 295
spectrum 296
spectrum 297
spectrum 298
spectrum 299
spectrum 300
n1 hl1
n2 hl1
n3 hl1
n4 hl1
n5 hl1
n6 hl1
n7 hl1
n8 hl1
n9 hl1
n10 hl1
n11 hl1
n12 hl1
n13 hl1
n1 hl2
n2 hl2
n3 hl2
n4 hl2
n5 hl2
n6 hl2
n7 hl2
n8 hl2
n1 hl3
n2 hl3
n3 hl3
n4 hl3
n5 hl3
n1 hl4
n2 hl4
n3 hl4
Predicted soil property

Training the model
(property + spectrum: step 1)

Prediction model
(spectrum only: step 2)

Figure 2.5: Design of the Artificial Neural Network, for training (top, step 1) and predicting (bottom, step 2).

(spectral data) were used to train the RF models, and spectral data (only, without soil data) applied into the models to assess their predictive capacity. The Random forest were used as the following R command `rf_model <- randomForest(form, data = trainData)`, where the `rf_model` is the Random Forest model; `randomForest` is the R function for Random Forest (from R package *randomForest*); `form` is the same *form* object as described above; and `data = trainData` is the same training data as described above.

To predict, it was applied was `fitted_rf <- predict(rf_model, newdata= validation_data)`. This function is similar to the one described above, although it used the `rf_model` (Random Forest model).

2.5.6 Partial Least Squares Regression (PLSR)

Partial Least Squares Regression (PLSR) is a multivariate regression technique widely considered for a large number of applications in several fields such as spectral analysis of food, agricultural products (LIU *et al.*, 2010), and spectral study of soils (ROSSEL *et al.*, 2005). PLSR establishes the relationship between highly collinear multi-dimensional predictor variables and the tested variable, thus, it selects the orthogonal factors to maximize the covariance between predictor and response variables (DANGAL *et al.*, 2019). The used R command is given by the expression `plsr_model <- plsr(formula= form, data= trainData, ncomp= 30)`, where the `plsr_model` is the PLSR model; `plsr` is the R function for PLSR (from R package *pls*); `formula= form` is the same *form* object as described above; `data = trainData` is the same training data as described above; and `ncomp= 30` is the the number of components of the model.

The predict function applied was `fitted_plsr <- predict(plsr_model, newdata = validation_data)`. This function is similar to the one described above, although it used the `plsr_model` (PLSR model).

2.5.7 Cubist (CB)

Cubist (CB) is a rule–based model used as an extension of model tree M5 (BASSER, 1992) that equates the need for accurate prediction with requirements for comprehensibility. The performance of CB has proved to be superior to other machine learning techniques, and it is simpler to understand since it is based on regression trees (NGUYEN *et al.*, 2019; Rulequest Research, 2019). The CB follows four steps (NGUYEN *et al.*, 2019): i) separation of data to grow a complete tree; ii) creation of a regression model at each node to prepare to pruning and prediction; iii) pruning the tree to evade overfitting problem; and iv) smoothing the tree to obfuscate the discontinuities limits caused by the splitting.

The used R command is given by the expression `cb_model <- cubist(x= trainData[, (spectral wavelengs)], y= trainData[ , current soil property], cubistControl(rules = 100, extrapolation = 15), committees = 1)`. The CB R algorithm implementation does not use a *form* object as the previous ones, instead it is necessary to make explicit the column on the data table where are the *x* and *y*; `cb_model` is the CB model; `cubist` is the R function for cubist (from R package *Cubist*); `x= trainData[, (spectral wavelengths)]` is the columns on the data table where the *spectral wavelengths* are placed; `y= trainData[ , current soil property]` is the column on the data table where the *current modeling soil property* are placed; `cubistControl (rules = 100, extrapolation = 15), committees = 1)` are rules: to define an explicit limit to the number of rules used, extrapolation: a number between 0 and 100: since Cubist uses linear models, predictions can be outside of the range seen training set. This parameter controls how much rule predictions are adjusted to be consistent with the training set; and committees: number of committee models (e.g. boosting iterations).

The predict function applied was `fitted_cb <- predict(cb_model, newdata= validation_data)`, this function is similar to the one described above, although it used the `cb_model` (CB model).

Since the main objective was to investigate the behavior of the spectral preprocessing, then compare the improvement of the preprocessing on different models, they were tested and applied with minimum model tuning, some of them close to the default parameters.

All the R commands and functions can be found in scripts format through the repositories of the Author, with respective data as well.

2.5.8 Models Performance Assessment, Cross-validation Approach

To assess the model's performance and avoid biased validation, the k-fold cross-validation was applied. The dataset was split into 10 folds (which means 10 groups) to perform cross validation, for this dataset each fold contains 30 samples. For every combination of preprocessing and modeling algorithm, on the 300 samples, each algorithm was applied 10 times. In other words, the training of 270 samples (9/10 of the total samples) and validating with 30 spare samples (1/10 of the total samples). This process was repeated until the algorithm predicted and validated all of the 300 samples (10/10) without repeating any sample for the validation (PEJOVIĆ *et al.*, 2018).

Across the 11 preprocessing, two (2) reduction of data dimensionality techniques, PCAL, raw data (without spectral preprocessing) for the 4 machine learning models (ANN, RF PLSR and CB), and 10 soil properties and the 10-split sampling (k-fold), a total of 6,000 models were created along the raw and treated data. Since each dataset is composed for 10 folds, the assessment (over the validated folds) was conducted as average over the 10 folds, then the 6,000 models were evaluate per group of 10, presenting 600 cross-validated results, which are presented in the Tables 2.3 (ranked the best preprocessing per soil property), 2.4 (ranked all preprocessing per soil property), and Figures 2.7 and 2.8 (associated with Table 2.3). All the 600 groups can be seen in Appendix A, Table S1 as supplementary material. Result of independent folds (for TC, P and K) are shown in the Table 2.5 (ranked per fold name).

To evaluate the performance of prediction models, the Root Mean Squared Error (RMSE), coefficient of determination (R^2) and the Ratio of Performance to Deviation (RPD) were calculated based on the average of the folds. All of the coefficients were calculated as an average across the folds. RPD is given by the ratio of standard deviation to the RMSECV (Root Mean Square Error of Cross-validation) or RMSEP (Root Mean Square Error of Prediction) between measured and predicted values (CHANG; LAIRD, 2002). Three classes of RPD are defined, where RPD>2 are the models that can predict well the soil property in analysis, RPD between 1.4 and 2 as an intermediate, and RPD<1.4 with no prediction ability (CHANG; LAIRD, 2002; GOMEZ *et al.*, 2008). The models were assessed essentially by the R^2, RMSE, bias and RPD.

2.5.9 Software

The software used to conduct the spectral readings was the ASD Rs3®.

The data processing and predictions were undertaken with R (R Core Team, 2019), with the packages: base R (R Core Team, 2019) and dplyr (WICKHAM *et al.*, 2019b) for data manipulation; rstudioapi (USHEY *et al.*, 2019) to automatically set working directory; caret (WING *et al.*, 2019) to find high correlated covariates; prospectr (STEVENS; RAMIREZ-LOPEZ, 2014) to visualize spectral data and apply preprocessing tools such as Savitzky-Golay and continuum removal; randomForest (BREIMAN *et al.*, 2018) for Random Forest, Cubist (KUHN; QUINLAN, 2018; Rulequest Research, 2019) for Cubist, pls (MEVIK *et al.*, 2019) for PLSR and neuralnet (FRITSCH *et al.*, 2019) for ANN predictor algorithms; stats from base R (R Core Team, 2019) for *predict* function; MASS (RIPLEY, 2019) for stepAIC application; ithir (MALONE, 2018) for metrics; RColorBrewer (NEUWIRTH, 2014), hexbin (CARR *et al.*, 2019), grid (MURRELL, 2014), and ggplot2 (WICKHAM *et al.*, 2019a) for graphs; DMwR (TORGO, 2013) to unscale the data after ANN; and stringr (WICKHAM, 2019) to access the results.

2.6 RESULTS

2.6.1 Laboratory Measured Soil Properties

The summary of statistics for the soil properties measured using conventional laboratory chemistry methods are presented in Table 2.2 and data distribution in Figure 2.6.

Table 2.2: Descriptive statistics for properties of soils sampled at the upper part of Itatiaia National Park, Rio de Janeiro State.

Soil property	Unit	N_s	Mean	SD	Median	Min	Max	Skew	Kurtosis	SE
Al^{3+}	$cmol_c\ dm^{-3}$	300	2.01	1.48	1.8	0	9.2	1.5	3.57	0.09
Ca^{2+}	$cmol_c\ dm^{-3}$	300	0.14	0.26	0.09	0	2.55	5.75	42.45	0.02
H^+	%	300	1.93	0.74	1.88	0.33	4.37	0.47	0.19	0.04
K^+	$cmol_c\ dm^{-3}$	300	0.14	0.15	0.08	0.01	1.26	3.06	14.64	0.01
Mg^{2+}	$cmol_c\ dm^{-3}$	300	0.5	0.3	0.42	0	1.67	1.53	2.26	0.02
N^+	%	300	0.36	0.35	0.25	0	1.85	1.43	1.93	0.02
Na^+	$cmol_c\ dm^{-3}$	300	0.03	0.05	0.03	0	0.8	10.95	149.19	0
P	ppm	300	7.33	8.47	4.81	0.19	97.51	5.03	43.28	0.49
pH	unitless	300	4.5	0.4	4.5	3.24	5.72	0.08	0.09	0.02
TC	%	300	5.9	5.62	3.99	0.24	29.48	1.38	1.71	0.32

Al: aluminum; Ca: calcium; H: hydrogen; K: potassium; Mg: magnesium; N: nitrogen; Na: sodium; P: phosphorus; TC: total carbon; N_s: number of samples; SD: standard deviation; SE: standard error.

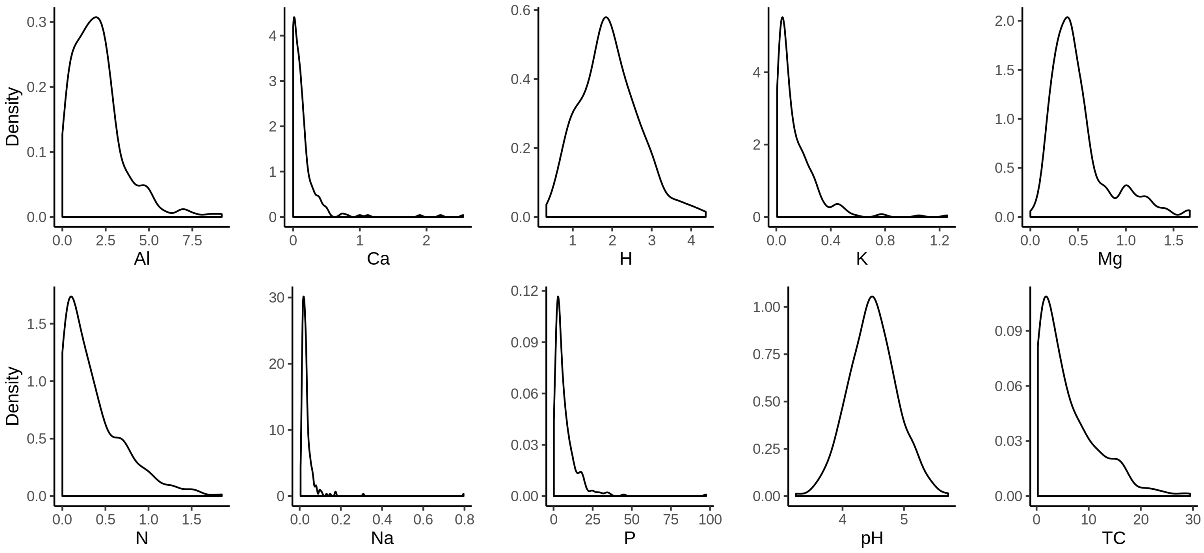


Figure 2.6: Density plot of properties of soils sampled at the upper part of Itatiaia National Park, Rio de Janeiro State.

Al: aluminum; Ca: calcium; H: hydrogen; K: potassium; Mg: magnesium; N: nitrogen; Na: sodium; P: phosphorus; TC: total carbon (Properties units according to Table 2.2).

The number of samples (N_s) was equal to 300 for all properties. Most of them deviate from normal distribution (Figure 2.6 and Skew from Table 2.2), except for H and pH values which presented skew and Kurtosis close to zero, and the pH is log-normally distributed.

2.6.2 Assessment of the Models

To evaluate the performance of predictive models, the Root Mean Squared Error (RMSE), coefficient of determination (R^2) and Ratio of Performance to Deviation (RPD) were calculated as the average of the folds. Of the 600 cross-validated groups, the best model associated with the best preprocessing, was the CB for TC with R^2 of 0.85, RPD of 2.87 (highest), followed by PLSR for N with R^2 of 0.82 and RPD of 2.65, and RF for Al with R^2 of 0.54 and RPD of 1.54 (Table 2.3). For contents of TC, Al, N, pH the RF presented the best association among the preprocessing tools giving higher ranking results 36 times, compared to 21 times for CB, 3 for PLSR, and none (0) for ANN (Table 2.4). For pH values the preprocessing significantly increased the performance bringing the R^2 from 0.096 to 0.36 (in comparison with no spectral treatment), followed by Al from 0.36 to 0.54. In general, the IRF4 and its association with SVG-1-2-11 improved the performance of the machine learning, followed by SVG-1-2-11 and their combination with NR (Table 2.4). Compared with others preprocessing, the widely used AB-log was midway favorable to TC with a slight improvement for N and less for Al values.

Table 2.3: Outstanding preprocessing with the associated models for each property of soils sampled at INP.

Preprocessing	Model	Soil property*	R^2	MSE	RMSE	bias	RPD
IRF4 + SVG-1-2-11 + NR 434	rf	Al	0.536	0.944	0.954	0.037	1.541
IRF4	cb	H	0.672	0.173	0.411	-0.034	1.817
SVG-1-2-11 + IRF4 + NR 434	rf	K	0.275	0.017	0.118	0.003	1.244
SVG-1-2-11	rf	Mg	0.194	0.074	0.267	0.014	1.148
IRF4 + NR 434	plsr	N	0.819	0.018	0.13	-0.005	2.649
SVG-1-2-11 + IRF4 + NR 434	rf	P	0.072	66.896	7.436	0.137	1.07
SVG-1-2-11	rf	pH	0.363	0.096	0.309	-0.005	1.286
IRF4	cb	TC	0.852	3.998	1.958	-0.044	2.867

*Ca and Na are not shown in the table due to the very poor results. The description of each preprocessing is according to Table 2.1; rf: random forest; cb: cubist; plsr: Partial Least Squares Regression; TC: total carbon; R^2: coefficient of determination; MSE: mean squared error; RMSE: root-mean-square error; RPD: ratio of performance to deviation. The coefficients units correspond to Table 2.2.

The reduction of dimensionality technique stepAIC, for pH and Al, in comparison with no preprocessed spectrum, weakly improved the prediction in most of the cases. Although considering that the model uses only 299 wavelengths (covariates) it still performed satisfactorily when considering the fact that it uses only 14% of all spectral data. As observed for the N values, in some cases, some preprocessing decreased the performance of the model, for example CR for TC (Table 2.4). The validation graphs of the properties Al, H, K, Mg, N, P, pH and TC are in Figures 2.7 and 2.8, according to the models and preprocessing of Table 2.3. The 600 cross-validated groups are presented as supplementary materials (Table S1).

Table 2.4: Selected cross-validated groups of the preprocessing with associated models for values of TC, N, Al and pH of soils from the INP.

Preprocessing	Model	Soil property	R^2	MSE	RMSE	bias	RPD
IRF4	cb	TC	0.852	3.998	1.958	-0.044	2.867
IRF4 + SVG-1-2-11	rf	TC	0.841	4.753	2.112	-0.038	2.65
IRF4 + SVG-1-2-11 + NR 434	rf	TC	0.84	4.749	2.113	-0.027	2.649

Table 2.4 – continued on the next page

Table 2.4 – continued from previous page

Preprocessing	Model	Soil property	R^2	MSE	RMSE	bias	RPD
SVG-1-2-11	rf	TC	0.836	4.718	2.123	0	2.627
SVG-1-2-9	rf	TC	0.833	4.84	2.147	0.014	2.604
SVG-1-2-11 + NR 434	rf	TC	0.831	4.854	2.143	0.024	2.632
AB-log	cb	TC	0.829	4.717	2.121	-0.14	2.652
SVG-1-2-11 + IRF4	rf	TC	0.826	5.047	2.194	-0.013	2.539
SVG-1-2-11 + IRF4 + NR 434	rf	TC	0.824	5.006	2.183	0.014	2.569
IRF4 + NR 434	plsr	TC	0.824	4.965	2.181	-0.055	2.559
no preprocessing	cb	TC	0.824	5.128	2.153	-0.059	2.667
CR	rf	TC	0.81	5.581	2.295	0.026	2.442
stepAIC	cb	TC	0.803	5.339	2.266	-0.052	2.466
PCAL	cb	TC	0.793	5.518	2.297	-0.148	2.338
RHCC	cb	TC	0.789	6.179	2.41	-0.166	2.338
IRF4 + NR 434	plsr	N	0.819	0.018	0.13	-0.005	2.649
IRF4 + SVG-1-2-11 + NR 434	rf	N	0.815	0.019	0.137	-0.002	2.466
IRF4 + SVG-1-2-11	rf	N	0.812	0.02	0.138	-0.002	2.446
AB-log	plsr	N	0.798	0.021	0.143	-0.003	2.37
SVG-1-2-11	rf	N	0.797	0.021	0.142	0.003	2.382
SVG-1-2-9	rf	N	0.791	0.021	0.143	0.003	2.37
SVG-1-2-11 + IRF4	rf	N	0.787	0.022	0.146	0	2.321
SVG-1-2-11 + NR 434	rf	N	0.786	0.021	0.144	0.005	2.375
SVG-1-2-11 + IRF4 + NR 434	rf	N	0.777	0.022	0.148	0.002	2.318
RHCC	cb	N	0.769	0.025	0.153	-0.008	2.265
CR	rf	N	0.755	0.026	0.157	0.004	2.166
no preprocessing	cb	N	0.743	0.028	0.162	-0.009	2.141
IRF4	cb	N	0.731	0.03	0.161	-0.005	2.232
stepAIC	cb	N	0.728	0.028	0.164	-0.009	2.085
PCAL	cb	N	0.701	0.031	0.173	-0.004	1.89
IRF4 + SVG-1-2-11 + NR 434	rf	Al	0.536	0.944	0.954	0.037	1.541
IRF4 + SVG-1-2-11	rf	Al	0.527	0.967	0.965	0.039	1.527
SVG-1-2-11 + NR 434	rf	Al	0.522	0.97	0.966	0.042	1.529
IRF4 + NR 434	cb	Al	0.517	0.935	0.961	0.003	1.514
SVG-1-2-11	rf	Al	0.513	0.987	0.974	0.049	1.52
SVG-1-2-9	rf	Al	0.511	0.981	0.973	0.045	1.516
SVG-1-2-11 + IRF4 + NR 434	rf	Al	0.506	1.028	0.992	0.032	1.485
SVG-1-2-11 + IRF4	rf	Al	0.487	1.071	1.012	0.034	1.459
AB-log	cb	Al	0.431	1.135	1.049	-0.079	1.405
stepAIC	cb	Al	0.423	1.152	1.062	-0.117	1.371
IRF4	cb	Al	0.419	1.078	1.027	-0.078	1.449
CR	rf	Al	0.388	1.225	1.097	0.062	1.322
no preprocessing	cb	Al	0.362	1.264	1.111	-0.072	1.319
RHCC	cb	Al	0.261	1.402	1.161	-0.076	1.295
PCAL	rf	Al	0.249	1.516	1.218	0.056	1.203
SVG-1-2-11	rf	pH	0.363	0.096	0.309	-0.005	1.286
SVG-1-2-11 + NR 434	rf	pH	0.352	0.098	0.312	-0.007	1.278

Table 2.4 – continued on the next page

Table 2.4 – continued from previous page

Preprocessing	Model	Soil property	R^2	MSE	RMSE	bias	RPD
SVG-1-2-9	rf	pH	0.352	0.098	0.312	-0.006	1.277
IRF4 + SVG-1-2-11 + NR 434	rf	pH	0.347	0.098	0.311	0	1.284
IRF4 + SVG-1-2-11	rf	pH	0.346	0.098	0.312	0	1.28
SVG-1-2-11 + IRF4 + NR 434	rf	pH	0.329	0.101	0.317	-0.005	1.255
SVG-1-2-11 + IRF4	rf	pH	0.322	0.102	0.319	-0.005	1.249
IRF4	cb	pH	0.22	0.117	0.34	-0.02	1.18
IRF4 + NR 434	cb	pH	0.21	0.118	0.342	-0.007	1.172
CR	rf	pH	0.18	0.124	0.351	0	1.133
AB-log	cb	pH	0.123	0.134	0.362	-0.014	1.115
PCAL	cb	pH	0.097	0.141	0.373	-0.015	1.087
stepAIC	rf	pH	0.097	0.137	0.369	0.004	1.082
no preprocessing	rf	pH	0.096	0.139	0.37	0.004	1.08
RHCC	rf	pH	0.071	0.141	0.374	0.003	1.067

The description of each preprocessing is on Table 2.1; rf: random Forest; cb: Cubist; plsr: Partial Least Squares Regression; TC: Total Carbon; R^2: coefficient of determination; MSE: Mean Squared Error; RMSE: Root Mean Square Error; RPD: Ratio of Performance to Deviation. The coefficients units correspond to Table 2.2.

The highest prediction with the best associated preprocessing for each soil property with the top ranked model is presented at Table 2.3. Following the same line, the best predicted properties (TC, N, AL, pH), leaving out K, Mg, and P, are shown in Table 2.4, and they are ranked according to the higher R^2 per property. The first line per property shows the most accurate result for the preprocessing with the associated model. All the preprocessing methods are displayed per property, although the model displayed is just the top one per preprocessing. P and K are displayed in Table 2.5 with the individual values of each fold. Table S1 is alphabetically organized according to: Preprocessing, Model, Soil property.

The common (and simplest) external validation was previously tested several times, with random selection of data: 70% for training and 30% for validation. Each random selection ended in a more or less homogeneous/heterogeneous groups of each analyzed soil properties, which turns in a very different prediction accuracy, for example varying the R^2 from 0.60 to 0.95 each time the algorithm performed the prediction. Even the k-fold cross-validation technique keeps the sample groups fixed into the folds, differently from external validation, this variability in the prediction accuracy can be observed within the folds of TC values (Table 2.5), it is observed that the R^2 validation coefficient ranges from 0.78 to 0.95. For the 10th fold the RPD is highest, and the best MSE and RMSE remains in fold 6. For P and K values, fold 8 has a huge difference in the coefficients. While the average of R^2 for P values is 0.072, in fold 8 it is 0.35, and for K we have an R^2 average of 0.275, while in fold 8 it is 0.674. The fold spectral behavior can be seen in Figure 2.9, (folds from 1 to 7 and 9 to 10 in orange, and 8 in blue).

As consequence of the work, a standard operational procedure was developed to use the spectroradiometer on the laboratory of radiometry (LabSpec) at the Federal Rural University of Rio de Janeiro. Which can be found on the repositories of the Author. The data and R code, can be found also on the same repositories.

Table 2.5: The coefficients within the 10 folds for TC, P and K.

Folds	R^2	MSE	RMSE	bias	RPD
TC.1	0.814	7.209	2.685	0.045	2.36
TC.2	0.899	3.947	1.987	-0.062	3.193
TC.3	0.894	3.203	1.79	-0.433	3.13
TC.4	0.781	7.54	2.746	0.055	2.174
TC.5	0.843	3.005	1.734	0.363	2.571
TC.6	0.9	2.486	1.577	0.047	3.218
TC.7	0.886	2.83	1.682	-0.155	3.018
TC.8	0.87	4.353	2.086	-0.26	2.822
TC.9	0.682	2.794	1.672	0.541	1.804
TC.10	0.946	2.617	1.618	-0.578	4.377
P.1	-0.163	79.717	8.928	-0.819	0.943
P.2	-0.207	37.404	6.116	1.477	0.926
P.3	0.222	34.459	5.87	0.736	1.153
P.4	0.18	56.343	7.506	0.204	1.123
P.5	0.272	24.237	4.923	0.177	1.192
P.6	-0.003	40.109	6.333	0.869	1.015
P.7	-0.023	291.659	17.078	-1.512	1.006
P.8	0.352	34.759	5.896	-0.659	1.264
P.9	-0.023	24.13	4.912	0.71	1.005
P.10	0.108	46.141	6.793	0.185	1.077
K.1	-0.034	0.058	0.24	-0.041	1
K.2	0.27	0.006	0.08	0.006	1.19
K.3	0.141	0.008	0.091	0.033	1.097
K.4	0.35	0.026	0.161	-0.007	1.261
K.5	0.163	0.025	0.158	-0.021	1.112
K.6	0.3	0.006	0.076	0.033	1.215
K.7	0.089	0.006	0.08	0.033	1.066
K.8	0.674	0.005	0.071	-0.001	1.783
K.9	0.576	0.004	0.063	0.024	1.561
K.10	0.225	0.026	0.16	-0.024	1.155

TC: Total Carbon; R^2: coefficient of determination; MSE: Mean Squared Error; RMSE: Root Mean Square Error; RPD: Ratio of Performance to Deviation. The coefficients units correspond to Table 2.2. Algorithms and preprocessing in this table: For TC, IRF4 with CB model; for P and K (SVG-1-2-11 + IRF4 + NR 434) with RF model.

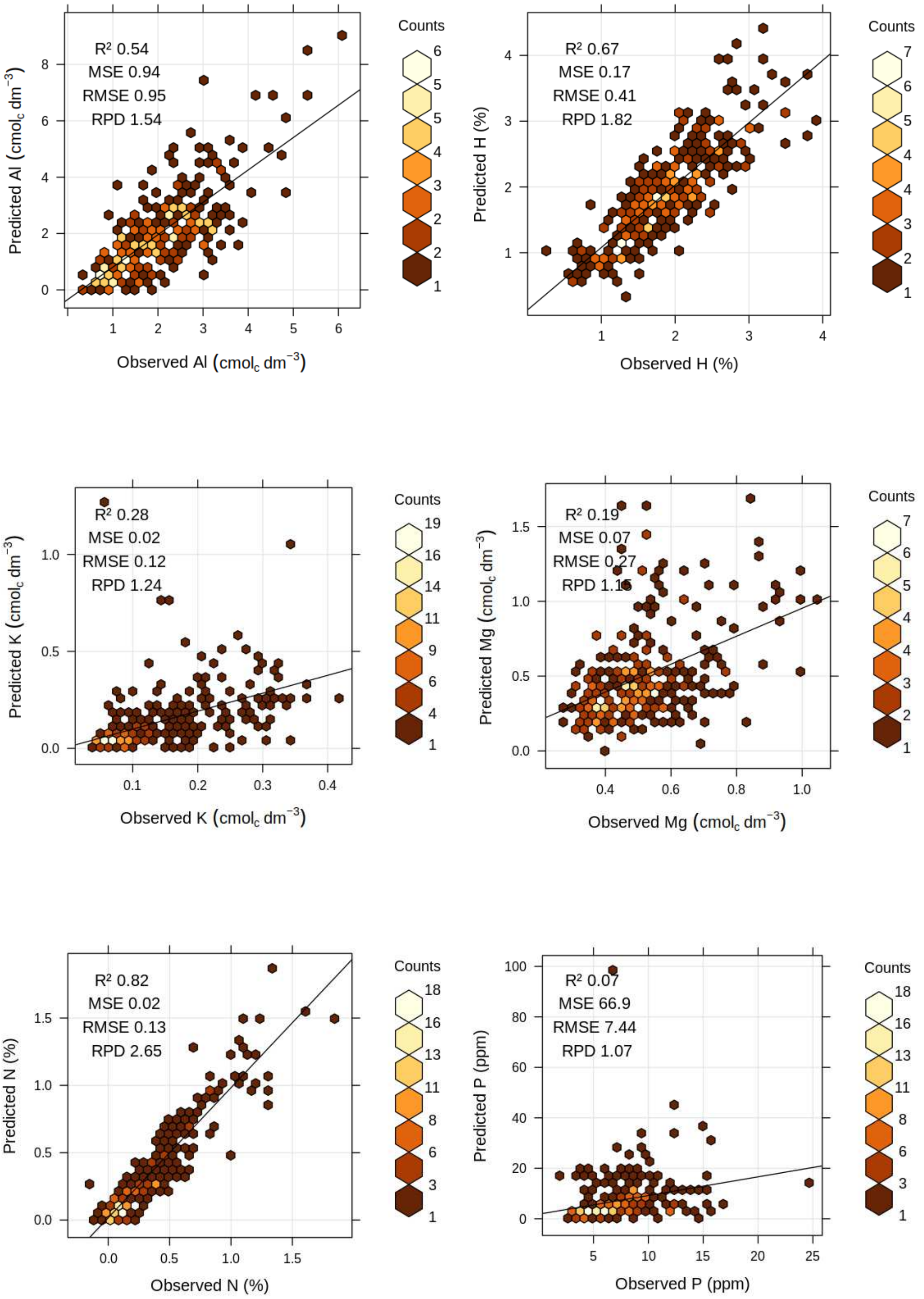


Figure 2.7: Prediction of Al, H, K, Mg, N and P. Displays only the top ranked predictions for each property along the best preprocessing with the best associated model, according to Table 2.3. Transversal line is the fitted correspondent model line, according to Table 2.3.

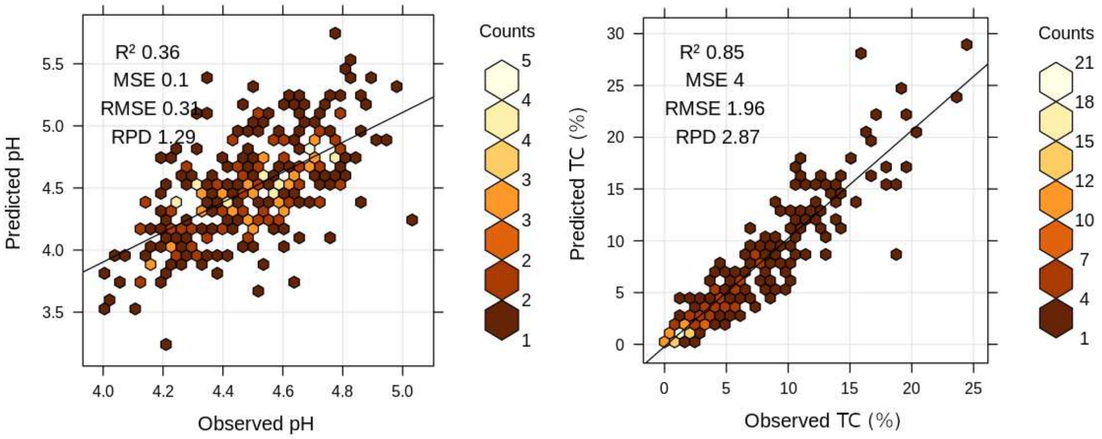


Figure 2.8: Prediction of pH and TC values. Displays only the top ranked predictions for each property along the best preprocessing with the best associated model, according to Table 2.3. Transversal line is the fitted correspondent model line, according to Table 2.3.

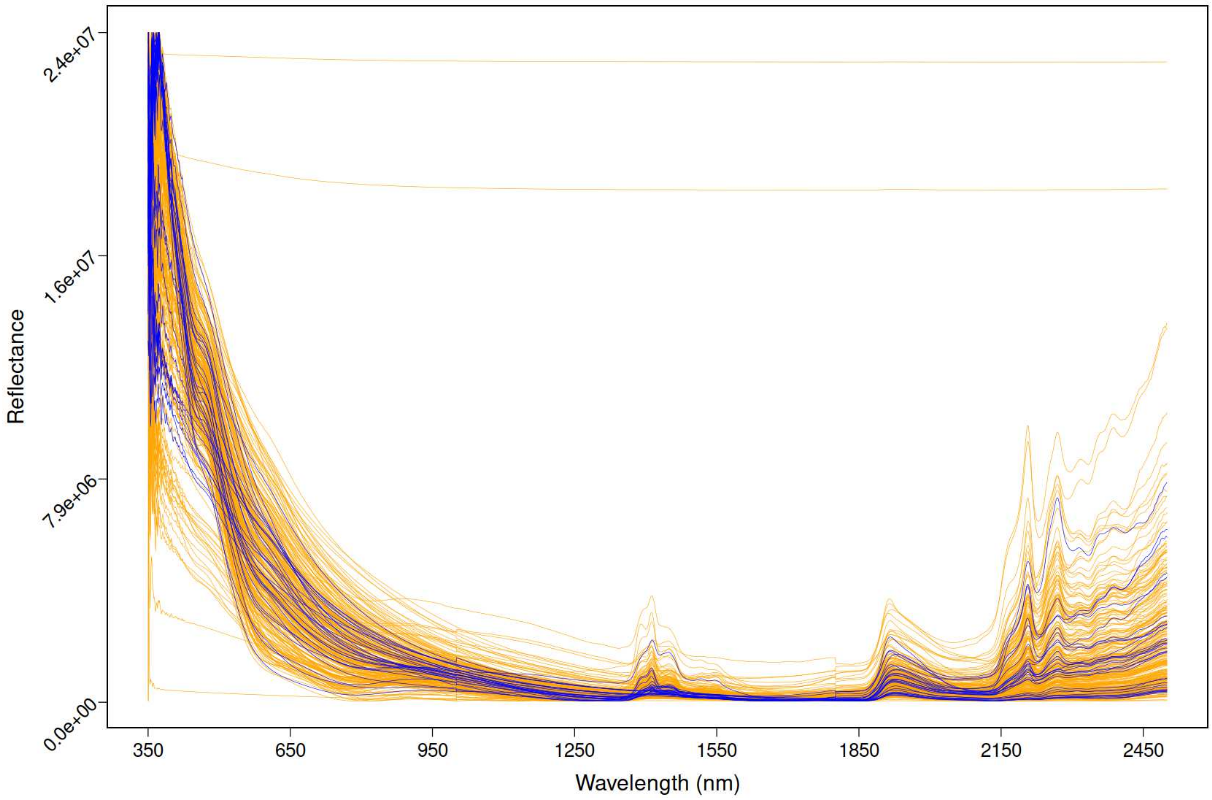


Figure 2.9: Spectral plots of soil samples from the upper part of Itatiaia National Park, with IRF4 preprocessing. Fold 8 is in blue.

2.7 DISCUSSION

The soil property that was best predicted was the TC, with R^2 0.85, RMSE 1.96, bias -0.04 and RPD of 2.87 (Figures 2.7 and 2.8). The other properties showed bias close to zero with the exception of values of P, which was 0.137. From the dimensionality reduction (Table 2.4), the RHCC, and especially stepAIC showed that the spectral resolution is not the main driver that improves the prediction of soil properties, which is in agreement with Gomez *et al.* (2008), but it still allowed the models to reach R^2 0.8, RMSE 2.26, bias -0.052 and RPD 2.47 for TC with CB model (Table 2.3). CR was similar to RF for TC giving R^2 0.81, RMSE 2.30, bias 0.026 and RPD of 2.44. In addition, the RHCC and stepAIC reduced the processing time, and thus machine power consumption. Differently from Dangal *et al.* (2019) the removal of outliers did not improve the prediction; however, this procedure may relevant when working with large databases. The PCAL with 5% removed still provided satisfactory results max of R^2 0.79 and RPD of 2.33.

The Sawitzky–Golay filter improved the prediction of the properties Al, K, Mg, P and pH (VASQUES *et al.*, 2008; KOPAČKOVÁ *et al.*, 2017), with similar results for ten out of the seventeen soil properties in central Amazon soils (PINHEIRO *et al.*, 2017) and for South Eastern Australia (TANG *et al.*, 2020). Our results show that the setting of SVG-1-2-11 provided higher coefficients in comparison with SVG-1-2-9. The application of IRF4 (combined or alone) benefited the models and increased the predictive capacity for 6 of the 8 predicted properties in comparison with the raw spectra (Table 2.3, 2.4 and S1), which are: Al, H, K, N, P, and TC. IRF4 alone and combined performed better than the commonly used preprocessing alone.

Close predictions of TC were obtained with ANN and with a bandwidth of 10 nm (DANIEL *et al.*, 2003), the bandwidth had a role in the prediction capacity, but it was not a key feature (GOMEZ *et al.*, 2008). Large spectral data and different ANN strategy could rise the ANN potential prediction (PADARIAN *et al.*, 2019). Testing a series of spectral preprocessing Dotto *et al.* (2018) concluded that CR had the highest performance associated to PLSR and RF, followed by SVG to RF.

Machine learning such as RF and CB showed good predictive capacity, and the results using ANN may be further improved with a larger dataset. A review of methods and results (ROSSEL *et al.*, 2005) show better prediction values with a variety of algorithms aside of ANN; where Chang *et al.* (2001) reached a R^2 of 0.89 using PLSR for TC in V-SWIR, which is a procedure widely applied in the literature. Using Convolution neural Networks, Padarian *et al.* (2019) pointed a possible limitation of PLSR. In this study we found, overall, that the Cubist model and Random Forest presented higher prediction capacity for TC. Then in accordance with Tang *et al.* (2020), along algorithms comparison, CB has also presented better performance than PLSR to predict soil properties.

Certain spectral preprocessing for certain properties can decrease the model performance compared with no preprocessing, as observed for N (IRF4 (alone), stepAIC, PCAL), and TC (CR, stepAIC, PCAL, RHCC) (Table 2.4). In the case of TC, because it absorbs relatively equally across V-SWIR wavelengths, it effectively impacts the continuum rather than specific absorption features. In this case, there is a physical reason to expect that CR might remove more signal than noise. The reduction of data dimensionality techniques (stepAIC, RHCC) generally decreases the performance of the models due to the loss of information from specific removed bands.

Despite a small variability in internal learning of machine algorithms on each run, the random selection data for calibration and validation of the models can produce from slightly to considerably different results among the folds, as observed for the properties TC, P, and K (Table

2.5). This effect can be dissipated with larger datasets. In terms of managing the data variability within the dataset, the k-fold cross validation presented consistent coefficients (Tables 2.3 and 2.4). Since they are given by the average of the coefficients across the folds, no matter how many times the process is repeated, the prediction confirmed by the validation coefficients is more stable and reliable than the simpler (commonly applied) external validation (e.g. 70/30). It is recommended to consider the spatial clusterization when selecting the random samples as Pejović *et al.* (2018), and it can lead to more stable results coefficients because the selected samples are spatially distributed and balanced.

The variation on prediction assessment coefficients can be observed inside the folds (Table 2.5), and for the reasons mentioned previously it is possible to observe that the models performed very differently for each fold. Supported by the visual assessment of this fold (Figure 2.9) where the spectral behavior of IRF4 folds from 1 to 7 and 9 to 10 in orange, with the fold 8 illuminated in blue, which showed less reflectance heterogeneity (from 950 to 2150 nm). This points to a better prediction for P and K inside fold 8 with the machine learning algorithms.

2.8 CONCLUSIONS

The soil properties, TC and N presented the best prediction capacity, followed by H and Al, with the use of V-SWIR and the prediction algorithms ANN, RF, PLSR, CB. The pH had the highest increment with the SVG-1-2-11 preprocessing comparing with raw spectra. As for the preprocessing, each soil property had the prediction potential increased by a specific spectral preprocessing. In this way, globally SVG plus associations increased the potential for prediction. For TC, IRF4 outperformed the commonly used preprocessing, including SVG, similarly, IRF4+NR434 outperformed SVG for N. The combination of both with NR also showed good responses from the algorithms. For some preprocessing of soil properties, such as CR for TC and IRF4 (alone) for N, the preprocessing decreased the potential for prediction in comparison with the non-treated (raw) spectra. Without spectral preprocessing (on raw spectral data), the CB model showed the good prediction capacity, followed by and RF.

IRF4 was the top ranked preprocessing for N values when combined with NR. And IRF4 was the best for TC (R^2 0.85), and also raises the prediction of H (R^2 0.67), both using CB. This is followed by PLSR for N and RF for Al. The algorithm most present among the higher predicted values was RF (5 out of 8). The IRF4 technique is first time introduced in spectroscopy, compared with the traditional preprocessing, it is very simple to apply with no tuning needed. It is recommended more studies to confirm the potential of IRF4. Alone and associated with other spectral preprocessing.

The k-fold cross validation provided consistent and reliable coefficient indicators. The spectral data heterogeneity within the folds tends to decrease with the larger datasets, raising the prediction capacity.

Considering that soil carbon is an indicator of soil health, quality and degradation, the results obtained from the applied spectral soil properties prediction techniques show potential of fast environmental assessment. In this sense the techniques can contribute for the Itatiaia National Park management and monitoring.

The V-SWIR techniques has potential to contribute predicting soil properties in other areas of Atlantic Forest and similar environments. The good correlation with V-SWIR indicates a potential for a fast monitoring of soil properties, such as organic carbon. They can be further associated to the environmental orbital remote sensing, especially in regions with limited access such as the INP.

3 CAPÍTULO II

SUBSURFACE HYPERSPECTRAL IMAGES, A STRONG INTEGRATION BETWEEN PROXIMAL SOIL SENSING AND DIGITAL MAPPING OF SOIL PROPERTIES

3.1 RESUMO

O Parque Nacional do Itatiaia (INP) está localizado ao sul do estado do Rio de Janeiro, na fronteira com os estados de Minas Gerais e São Paulo, na região sudeste do Brasil. Sendo uma unidade de conservação, o INP é área de referência para estudos ambientais no Bioma Floresta Atlântica. Pelo seu relevo montanhoso, com acesso difícil e trilhas limitadas, além de afloramentos de rochas dominantes na parte alta do INP, o uso de ferramentas de Sensoriamento Remoto (RS, do inglês *Remote Sensing*) pode auxiliar o Mapeamento Digital de Solo (DSM, do inglês *Digital Soil Mapping*). Técnicas de RS que envolvem comprimentos de onda visíveis, infravermelho próximo e infravermelho de onda curta (Vis-NIR-SWIR ou simplesmente V-SWIR) são aplicáveis para a predição espacial de propriedades do solo. O objetivo do estudo foi combinar técnicas de RS, como o Sensoriamento Remoto Proximal (PSS, do inglês *Proximal Soil Sensing*), ao DSM para predizer espacialmente o conteúdo de Carbono Total (TC) dos solos no INP. Foram utilizadas três cenas de imagens hiperespectrais do sensor CHRIS (*Compact High Resolution Imaging Spectrometer*) embarcado no satélite PROBA (*Project for On Board Autonomy*). As imagens de 62 bandas (411 a 997 nm, referente ao meio da primeira e última bandas, respectivamente) foram corrigidas quanto a ruídos, *striping*, correções geométricas e atmosféricas. De posse dessas imagens CHRIS, associadas a covariáveis de relevo e imagens RapidEye, foi feita a predição de TC, atingindo coeficiente de determinação (R^2) de 0,33; enquanto que excluindo imagens CHRIS foi obtido R^2 de 0,32. Essas imagens foram combinadas com os espectros proximais obtidos de material da primeira camada do solo, em 84 pontos amostrados na parte alta do INP. Desta forma, foi possível produzir imagem da subsuperfície do solo, em outras palavras, uma imagem hiperespectral de subsuperfície. Com essas imagens em formato *raster*, a predição de TC apresentou R^2 de 0,58, ou seja, incremento de 75% na predição quando comparada ao DSM sem esta etapa. Essa técnica eliminou os efeitos da interferência atmosférica e da vegetação na reflectância do solo. Esses resultados trazem a primeira integração efetiva entre PSS e DSM, nomeado pelo autor de Mapeamento Hiperespectral de Solos (HSM, em inglês *Hyperspectral Soil Mapping*). Essa correlação entre as técnicas V-SWIR, imagens hiperespectrais e propriedades do solo, pode ser amplamente aplicada no mapeamento das propriedades do solo, sendo útil para fins agrícolas e para monitoramento ambiental remoto. Este último ainda mais relevante em áreas como o INP.

Palavras-chave: PSS+DSM. Métodos de Mapeamento. Espectrorradiometria.

3.2 ABSTRACT

Itatiaia National Park (INP) is located in the south of the state of Rio de Janeiro on the border with the states of Minas Gerais and São Paulo, in the southeastern region of Brazil. As a conservation unit, the INP is a reference area for environmental studies in the Atlantic Forest Biome. Due to the mountainous relief, with difficult access and limited trails, besides rock outcrops, manly in the upper part of INP, the use of Remote Sensing (RS) tools can assist the Digital Soil Mapping (DSM). Such RS techniques involving the visible, near-infrared and short-wave infrared (Vis-NIR-SWIR, or simply V-SWIR) wavelengths, are applicable for spatial prediction of soil properties. The objective of the study was to combine RS techniques, such as Proximal Soil Sensing (PSS), and DSM to spatially predict the content of total carbon (TC) in the soils of INP. Three scenes of hyperspectral images from the space platform Project for On Board Autonomy (PROBA), and the sensor Compact High Resolution Imaging Spectrometer (CHRIS) were used. The image of 62 bands (411 to 997 nm, referring to the middle of the first and last bands, respectively) were corrected for noise, striping, geometric and atmospheric corrections. TC was predicted using the CHRIS images, associated with relief covariates and RapidEye images, and the coefficient of determination (R^2) of 0.33; while without CHRIS images a R^2 of 0.32 was obtained. These images were combined with the proximal spectra obtained from soil samples taken from the first layer of 84 points in the upper part of INP. In this way, it was possible to produce an image of the subsurface soil, in another words, a subsurface hyperspectral image. With these raster images, it was possible to obtain a prediction of TC with R^2 of 0.58, which is a gain of 75% in comparison with the Digital Soil Mapping without this step. This technique eliminated the interference of atmosphere and vegetation on soil reflectance. These results bring the first strong integration between PSS and DSM, named by the author as Hyperspectral Soil Mapping (HSM). This correlation between V-SWIR techniques, hyperspectral images and soil properties, can be widely applied for mapping soil properties, being useful to agricultural purposes and remote environmental monitoring. This last even more relevant in areas such as the INP.

Keywords: PSS+DSM. Mapping Methods. Spectroradiometry.

3.3 INTRODUCTION

The land coverage, mainly by various types of vegetation, is one of the site characteristics that can diminish the capacity of the models to predict soil properties. The Digital Soil Mapping (DSM) techniques apply various methods and procedures enabling to minimize this effect through the use of covariates like the Digital Elevation Model (DEM) with its terrain derivations (for example slope and northernness), geology, and geomorphology, and, spectral reflectance bands from satellite imagery or Proximal Soil Sensing (PSS). Thus, researches and developments to find how new covariates can improve the DSM are very important (MCBRATNEY *et al.*, 2003).

One definition for DSM is "The creation and population of spatial soil information systems by numerical models inferring the spatial and temporal variations of soil types and soil properties from soil observation and knowledge from related environmental variables" (LAGACHERIE *et al.*, 2007). In other words, DSM applies geospatial techniques through mathematical models with computational tools to build a spatial relation among the list of chosen (spatial) covariates to generate a map of soil classes or attribute(s). The role and experience of the pedologist is extremely important to go along with the model, in order to interpret and deliver realistic results. It is noteworthy that, in the DSM, ∼71 covariates can be chosen among the pool, from those, 48 are derived from DEM (COELHO *et al.*, 2019).

The list of covariates is usually linked with soil-forming factors such as climate, organisms, relief, parent material, and time, as in the soil function proposed by Jenny (1941). Although the DSM went beyond and uses the *scorpan* model, which is a Jenny-like formulation for quantitative descriptions of relationships between soil and other spatially referenced factors. It is represented as – $S_{c,a} = \{s, c, o, r, p, a, n\}$, where the soil class ($S_c$) or soil attribute ($S_a$) are expressed by: *s*: soil, other properties of the soil at a point; *c*: climate, climatic properties of the environment at a point; *o*: organisms, vegetation or fauna or human activity; *r*: topography, landscape attributes; *p*: parent material, lithology; *a*: age, the time factor; *n*: space, spatial position (MCBRATNEY *et al.*, 2003).

The goal of a soil survey (data where maps are deriving from) is to subset a heterogeneous area in (more) homogeneous sections, with less variability according to soil attributes and to distinguish soil classes (IBGE, 2015). Itatiaia National Park (INP) is characterized by mountainous relief and the upper plateau has land cover classes ranging from dense forest to sparse vegetation, bare soil and rock outcrops, stratified according to the altitude. Therefore, the non-homogeneous features of INP bring more variables to account for soil mapping, and this complexity makes the INP to be an appropriate site to test new methods and to create covariates that may improve soil mapping.

The Total Carbon (TC) content can be used as a proxy to assess the soil health and quality (LAL, 2016) and it is a relevant variable for environmental monitoring. This soil property also has one of the best correlations in studies using Remote Sensing (RS) techniques (GALVÃO; VITORELLO, 1998). The INP has restricted access within the upper part of the park, and the locomotion is mainly by uneven and long trails. Due to local climate and elevation, the soils have very high TC for tropical standards, reaching up to values of 29,5% of organic carbon (COSTA *et al.*, 2020). Thus, the RS techniques may provide the appropriate tools to assist in DSM for spatial prediction of soil properties, such as TC. Techniques involving the visible, near-infrared and short-wave infrared (Vis-NIR-SWIR, or simply V-SWIR) wavelengths are also proved methods for soil analysis (BOWERS; HANKS, 1964; XIE; LI, 2016; ADELINE *et al.*, 2017; KOPAČKOVÁ *et al.*, 2017), and it is projected that under some circumstances they may even replace the wet chemistry analysis.

In the range of V-SWIR, the PSS (few centimeters from reading surface) has a very high spectral resolution and broader spectral range in comparison with orbital images. In the laboratory, with no external interference, the readings are very consistent, but there is a lack of spatial coverage. In the field, each point has to be measured individually. On the other hand, the spectral part often used for DSM has the entire area structured by pixels, usually called raster image or orbital image (hundreds of kilometers from the earth surface). Although, the satellite image lacks where PSS has strength, spectral resolution, and very low ambient interferences. The combination of PSS with DSM opens a possibility for a new set of covariates for predicting soil properties, mainly when using the so called - Subsurface Hyperspectral image, with its numerous bands.

To choose the best sensor for a given project it is important to compare the specifications of the different instruments. The orbital sensor Compact High Resolution Imager (CHRIS) from the space platform Project for On Board Autonomy (PROBA) satellite has 36 meters of spatial resolution, 62 bands with spectral range from 411 to 997 nm (Vis and part of NIR, at the middle of the first and last bands respectively). The wavelengths (middle) of each 62 CHRIS bands can be found in the item 3.4.7. This equipment is in the category of Hyperspectral sensor, being currently a good option to develop the combination of PSS with DSM due to its spectral resolution. The Proximal Sensor ASD FieldSpec 4, has no spatial resolution, then one point can be used as one pixel at each scan with very high spectral resolution. In this regard comes the satellite image strength, many readings at once, spatial resolution expressed by the pixels. The spectrordiometer (proximal sensor) has a spectral range from 350 to 2500 nm (V-SWIR) with smalls bandwidth 1.4 nm @ 350-1000 nm and 1.1 nm @ 1001-2500 nm.

Many studies have used PSS to enhance the prediction capacity through calibration, for example, a temporal series from Landsat 5 was created by the extraction of the bare soil pixels across the time to compose a synthetic image (DEMATTÊ *et al.*, 2018; GALLO *et al.*, 2018; MENDES *et al.*, 2019; PADILHA *et al.*, 2020). As a possible covariate the *bare ground images* can be used (WADOUX *et al.*, 2020). In similar approach *spectral-temporal response surface* was applied by Lamichhane *et al.* (2019).

The motivation of this study originates from the need to improve the spatial prediction of soil properties and thus to enhance DSM through the development of the Hyperspectral Soil Mapping (HSM), especially in regions such as INP with difficult access and, consequently, hard field sampling work for soil survey. Summarizing, there is room to test the insertion of PSS in DSM.

The hypothesis of this study are: i) A Hyperspectral image can produce better prediction of TC than a Multispectral one; ii) The spectral equalization and shadow treatment of a Hyperspectral image produces better result than without this steps; iii) The combination of PSS and DSM can enhance the spatial prediction of soil properties in comparison to the conventional DSM; iv) The combination of PSS and DSM to predict of soil properties, using a Hyperspectral image provides better result than Multispectral one; and v) Using the entire spectrum of wavelengths of PSS (350 to 2500 nm) gives a better result than the CHRIS bands wavelength range (411 to 997 nm).

The aim of this study was to evaluate the combination of the PSS and DSM approaches to improve the spatial prediction of Total Carbon of soils at the upper part of the INP. In practical terms, to combine the CHRIS Hyperspectral and RapidEye Multispectral images with laboratory spectral reading from PSS to generate a new set of covariates, such as Subsurface Hyperspectral image, and use it to refine the conventional DSM.

3.4 MATERIAL AND METHODS

3.4.1 Site Location, Description

Itatiaia National Park (INP) is located in the south of the state of Rio de Janeiro on the boundaries with states of Minas Gerais and São Paulo, in the southeastern region of Brazil. The highest peak, Agulhas Negras, has 2,791.55 msnm. The study was carried out in the so-called upper plateau of the INP, which is defined above the 2000 msnm elevation. A detailed description of the study area, soil and spatial sampling, analysis, and data handling are presented in the Chapter I, item 2.5 Material and Methods.

The soil sampling points, selected with cHLS method, are shown in Figure 3.1 as yellow dots, and the tracks as brow lines, greenish and brownish areas represents the lower and higher elevation respectively. As mentioned in the Chapter I, due to environmental factors, the TC increases with the elevation, as illustrated by the amount of TC (at each sampling point) as a function of elevation (Figure 3.2).

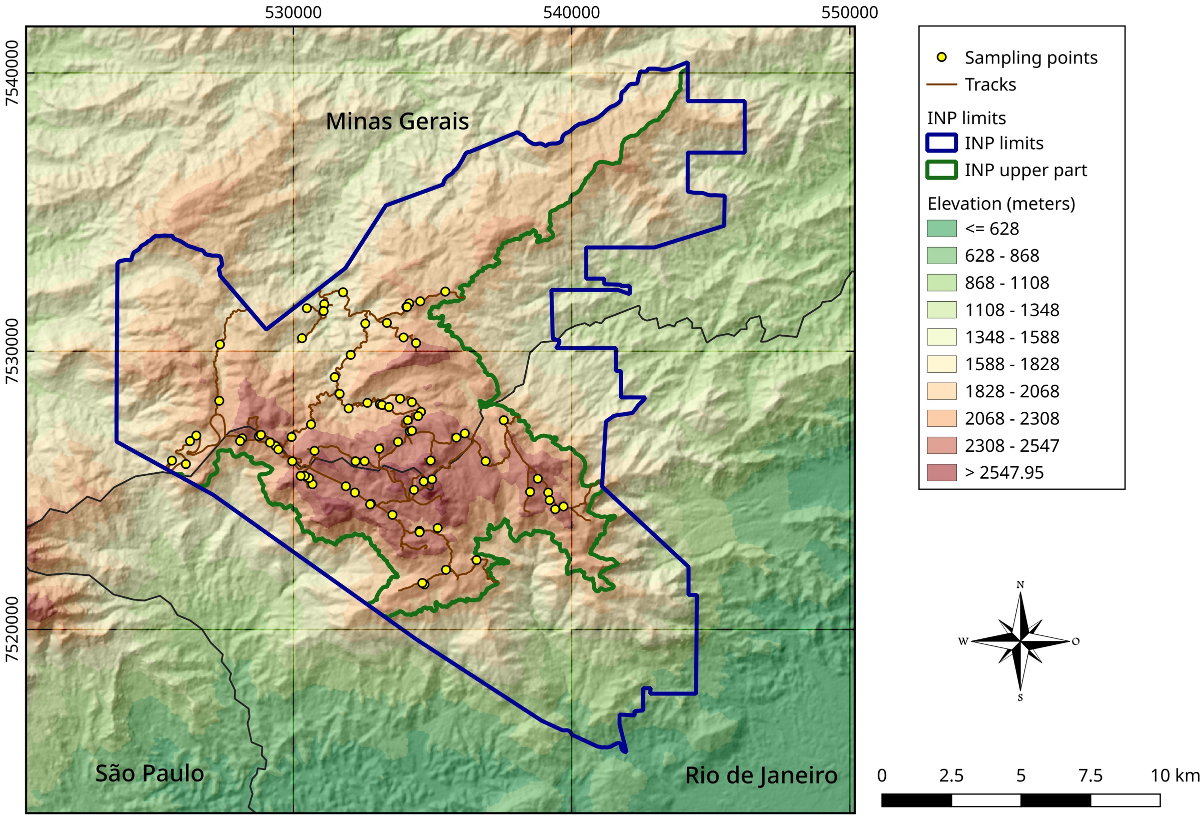


Figure 3.1: Itatiaia National Park (blue). The INP plateau (green line) at central and northwest areas, Mauá and lower part of INP areas at east and south respectively. Source of area delimitation lines: IBGE (2010), INP managers including Tomzhinski *et al.* (2012).

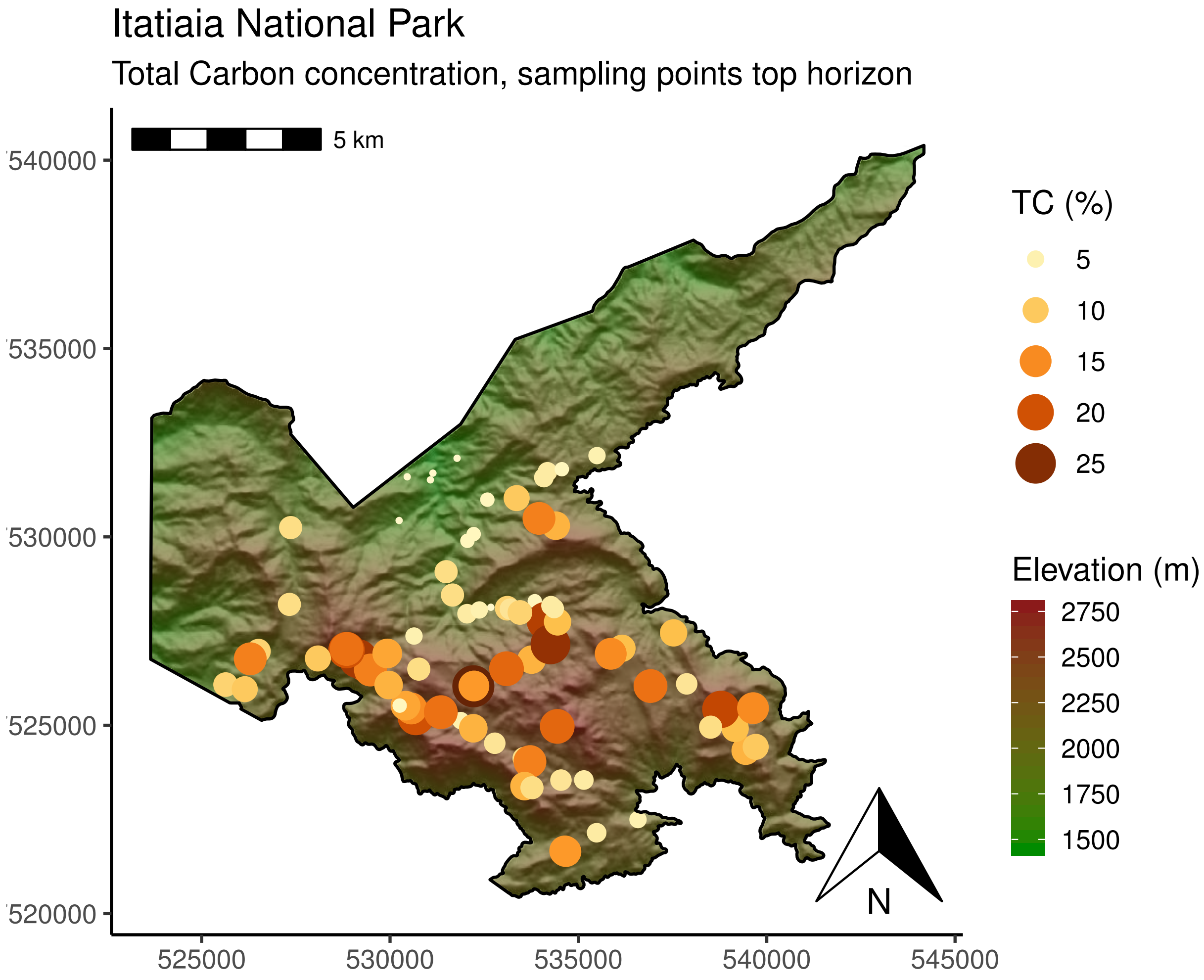


Figure 3.2: Distribution of Total Carbon in the soil according to the elevation in the upper part of INP, Rio de Janeiro state.

3.4.2 Sampling Spatial Dependence

In any survey, the quantity of samples is always a big debate, and there is a thin line between cost and effect. Too many sampling points, too expensive; too little, poor modeling results. Since INP upper part has hard access it was applied the cLHS (as detailed in the previous chapter) to define 84 profiles or sampling points, which made the survey costly in terms of energy and resources. From this collection, four samples were eliminated due to uncertainty on data. To keep confident on the spatial analyses spatial dependence analyses with a semivariogram was performed.

The semivariogram is a graph that express the increase of the sample's variance to the extent of the sampling locations (in distance). When the curve reaches the max variance it is called sill, which is the variance *a priori* of phenomenon. The further distance to get the max variance is the range. The sill and range are characterized by the "moment" when the curve turns horizontal. With the semivariogram is possible to understand the spatial dependence (spatial continuity) of the field samples (TZIACHRIS *et al.*, 2017).

Theoretically, on very short distance, the sampling variance tend to be zero, but in practical applications it rarely happens. So, in the semivariogram there is a discontinuity in the origin, called nugget effect. To explain the idea of nugget effect: in terms of TC, two soil samples very close usually have slightly different values, giving a smaller variance in comparison with other samples. The variance increases with the distance, when the variance reaches the sill, and the curve became horizontal, from that distance, which is called range, there is no spatial dependence anymore. With the semivariogram is adjusted a continuous function which allows to know the variance at any distance. Samples within the range have a certain statistical similarity.

The INP semivariogram (Figure 3.3) has nugget value of 0.72, sill 31.4, and range of 745 m. The samples were collected relatively close to each other, alongside the trail in a 100 m buffer area, which agrees to the low value of the nugget, confirms that near samples spots have lower variance. The sill is indeed close to the statistical variance which is 33.21, and the range tells that until 745 m distance there is spatial dependence within the TC soil samples.

The semivariogram was calculated with `variogram()` function and the model adopted was the spherical type, adjusted with `fit.variogram()` function to automatically find the best fit, both from R package `gstat`.

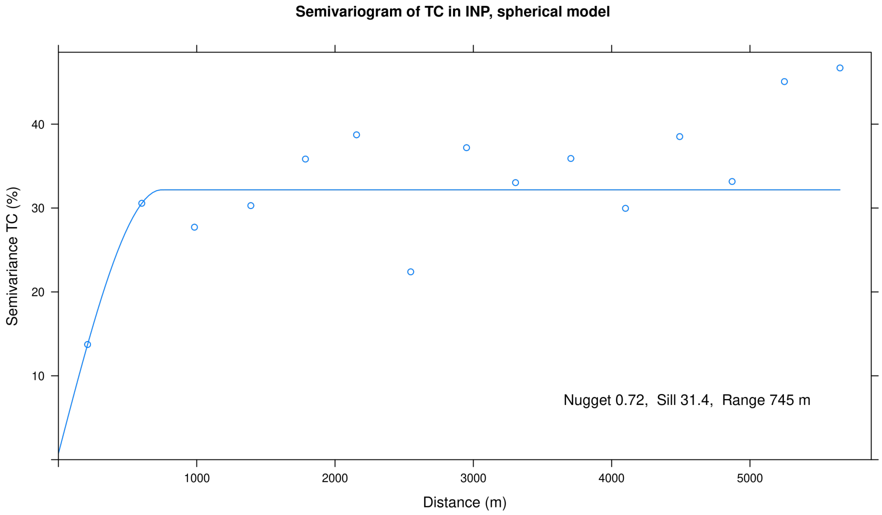


Figure 3.3: Semivariogram of Total Carbon (TC) in soils from upper part of INP with the spherical model.

3.4.3 The Multispectral Images (RapidEye)

The RapidEye constellation is a group of 5 satellites launched in August 2008 into a formation within the same orbital plane. They carry identical sensors calibrated to a common standard, in this way images from one satellite will be equivalent to any other four. It collects imagery in the blue (440-510 nm), green (520-590 nm), red (630-685 nm), the called Red edge (690-730 nm) and NIR (Near Infrared) (760-850 nm). The tile grid defines 24 by 24 km tiles, with a 1 km overlap, resulting in 25 by 25 km tiles (CAMPBELL; WYNNE, 2011). With the constellation, the temporal resolution is one day.

The RapidEye images from 2011/07/02 and 2011/08/16 were used, they have a 12-bit radiometric resolution, 6.5 m spatial resolution, and were orthorectified to 5 m spatial resolution (RapidEye, 2012). The images were atmospherically corrected using the model 6s adapted (ANTUNES *et al.*, 2012) an adaptation of the 6s model (VERMOTE *et al.*, 1997), as described in Costa *et al.* (2020). The RapidEye images were supplied through a license to Soils Department, UFRRJ, by the Brazilian Ministry of the Environment.

3.4.4 The Hyperspectral Images (CHRIS PROBA)

3.4.4.1 Acquisition, conventional treatment

The satellite PROBA was launch on 22 October 2001 (BEGIEBING; BACH, 2004), carrying the Hyperspectral sensor CHRIS onboard. This sensor produces a scene with 5 images across the zenith angles +55°, +36°, 0° (nadir), -36°, -55° (forwards, nadir and backwards). The platform has also the ability to shoot images of track ranging between 3.1 and 3.9 km wide (CUTTER; KELLAR-BLAND, 2008). The wavelengths of each 62 CHRIS bands can be found in the item 3.4.7 Subsurfacing Image Process, the CHRIS Operation further in the text.

During the dry season, in 2017/06/08, 2017/07/13, and 2018/08/11, 3 scenes were acquired on demand from the European Spatial Agency (ESA), through a project submission for a window-time on the dry season (not specific date). The first two had five view angles images and the last four. The images with fewer clouds and best cover of the sampling points were chosen from the second and third scenes, corresponding to the images: 2017/07/13 (code 3A60_41, 0°) and (code 3A61_41, +36°), together with 2018/08/11 (code 4CDF_41, 0°). Accounting the areas with vegetation as a nonlambertian target (as the natural targets (VERMOTE *et al.*, 1997)), we accept the trade-off to merge different angles scenes in favor to be able to cover all sampling points with the Hyperspectral images, when the different angle view (+36°) is the probable cause of lower reflectance of the image (3A61) (Figure 3.4). Even covering all sampling points, the scenes do not cover the entire INP plateau, thus, the mapping is restricted to this covering area when this scenes are used. The list of acquired images can be found in Appendix B, Table S2 as supplementary material.

Several image treatments were conducted, such as noise reduction (striping and odd pixels), atmospheric and geometric correction. The noise reduction was done with HDFClean V2 software (CUTTER, 2006), the atmospheric and geometric correction were performed within Beam/ESA software workflow menu commands (Brockmann, 2014) with a CHRIS module dedicated software which used the image metadata to accomplish the process in an automated form. The atmospheric correction algorithm is an adaptation of the MODerate resolution TRANsmittance (MODTRAN4), which deals with simultaneous aerosol and water vapor retrieval. The module also converts the images from Top-of-Atmosphere radiance to surface reflectance images (GUANTER *et al.*, 2008).

To perform the corrections, extra metadata was imported from external source provided by ESA. It was not possible to finish the last step of the process in the image 4CDF (Figure 3.4), the conversion from radiance to surface reflectance, due internal error of the software linked with external metadata. Still we were able to use the image by dividing each band for 10000 and making additional adjusts with R software, as described in the item 3.4.4.2 Scenes reflectance intensity equalization.

After these steps, the selected images needed georreference adjusts, done with the Quantum GIS software (QGIS Development Team, 2019), having a minimum of 4 ground control points and using as reference the RapidEye images (Table 3.1). Three images sufficiently covered the site over the sampled points. The three images still had different pixels sizes, which were adjusted in R software (R Core Team, 2019) with *aggregate*, *disaggregate* and *resample* from R *raster* package (HIJMANS, 2019), resulting in the mosaic image of Figure 3.4. From this point onward, all the processing was conducted in R software, with packages described in the Chapter I, item 2.5.9 Software.

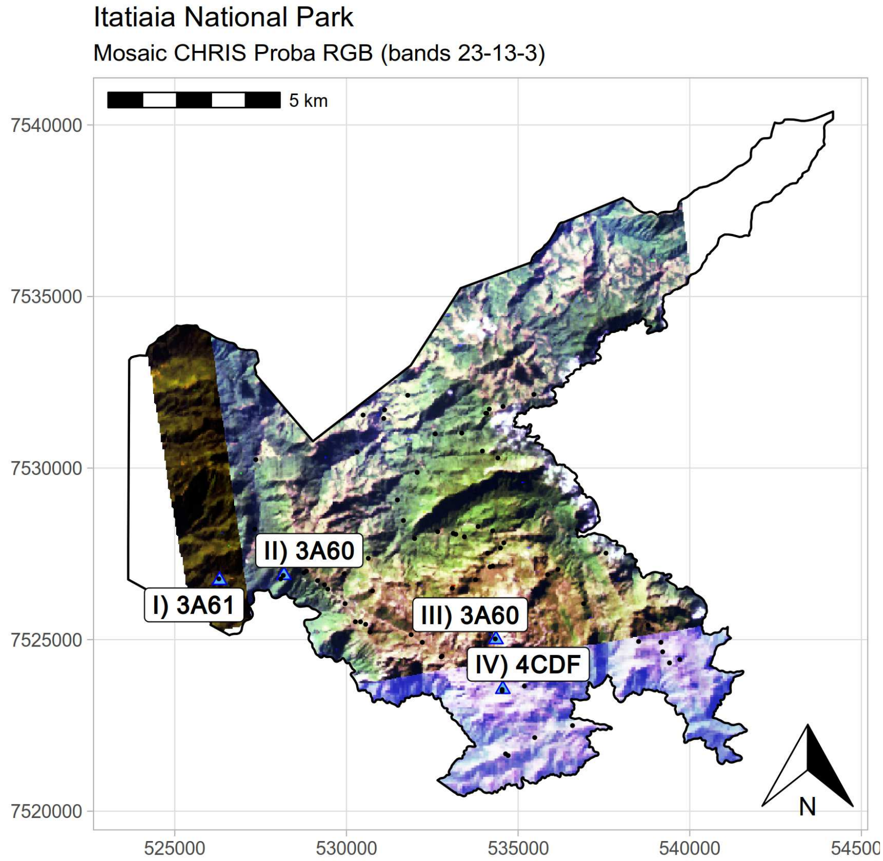


Figure 3.4: Mosaic of CHRIS PROBA images geometric and atmospheric corrected (3A60 center, 3A61 left, 4CDF bottom) of the INP plateau, Rio de Janeiro State.

3.4.4.2 Scenes reflectance intensity equalization

The Hyperspectral image presented differences of reflectance (Figure 3.4), also present in NIR, not only in RGB bands (Figure 3.6a). With tests of TC prediction accomplished, the differences were still present in the final map of TC (Figure 3.16a, clear in comparison with Figure 3.17a). This poor result motivated the spectral harmonization of the scenes. The reflectance of each scene had different intensity (Figure 3.4), thus the scenes were equalized to generate a smooth and uniform image (Figure 3.5). Aside from the ground variations, the reflectance over the mosaic area, between two scenes, has to be near constant. The images were adjusted for each band with raster operations, taking as reference the central image (3A60), which covers the major area and presents a better RGB plot (bands 23, 13 and 3, corresponding to 651, 551 and 452 nm) (Figure 3.4). As reference, the image (3A60, center) was not modified.

The raster operations started with extraction of pixels values from two draw lines, one of each side of each scene and very close to the mosaic borders (of each of the three scenes). This process was executed for all 62 bands in batch. For each band, the mean of these extracted (border) pixels from each scene were subtract to generate an index. Adding this index for each band we equalized the reflectance intensity among the scenes, resulting in a homogeneous mosaic (Figure 3.5).

To compare the treated images, four pixels (spatial points represented in Figure 3.4 as blue triangles) were chosen very close to the sampling points, in a bright area. The pixels' values were extracted from the mosaicked image before and after the correction, as seen in Figures 3.4

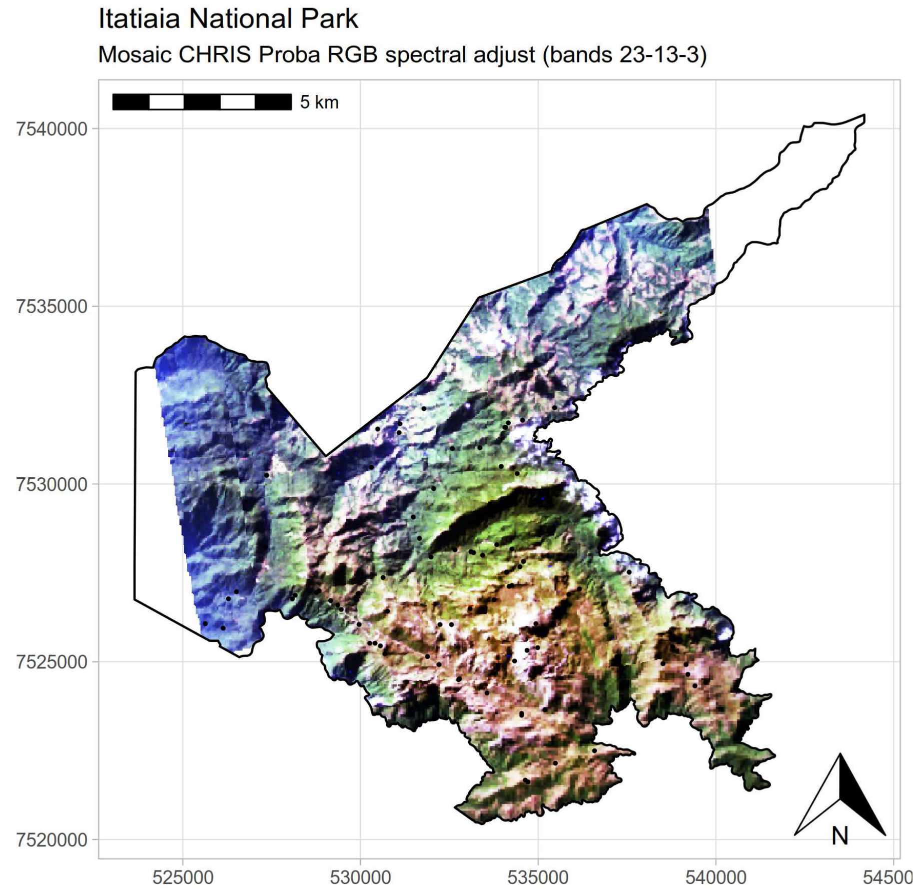


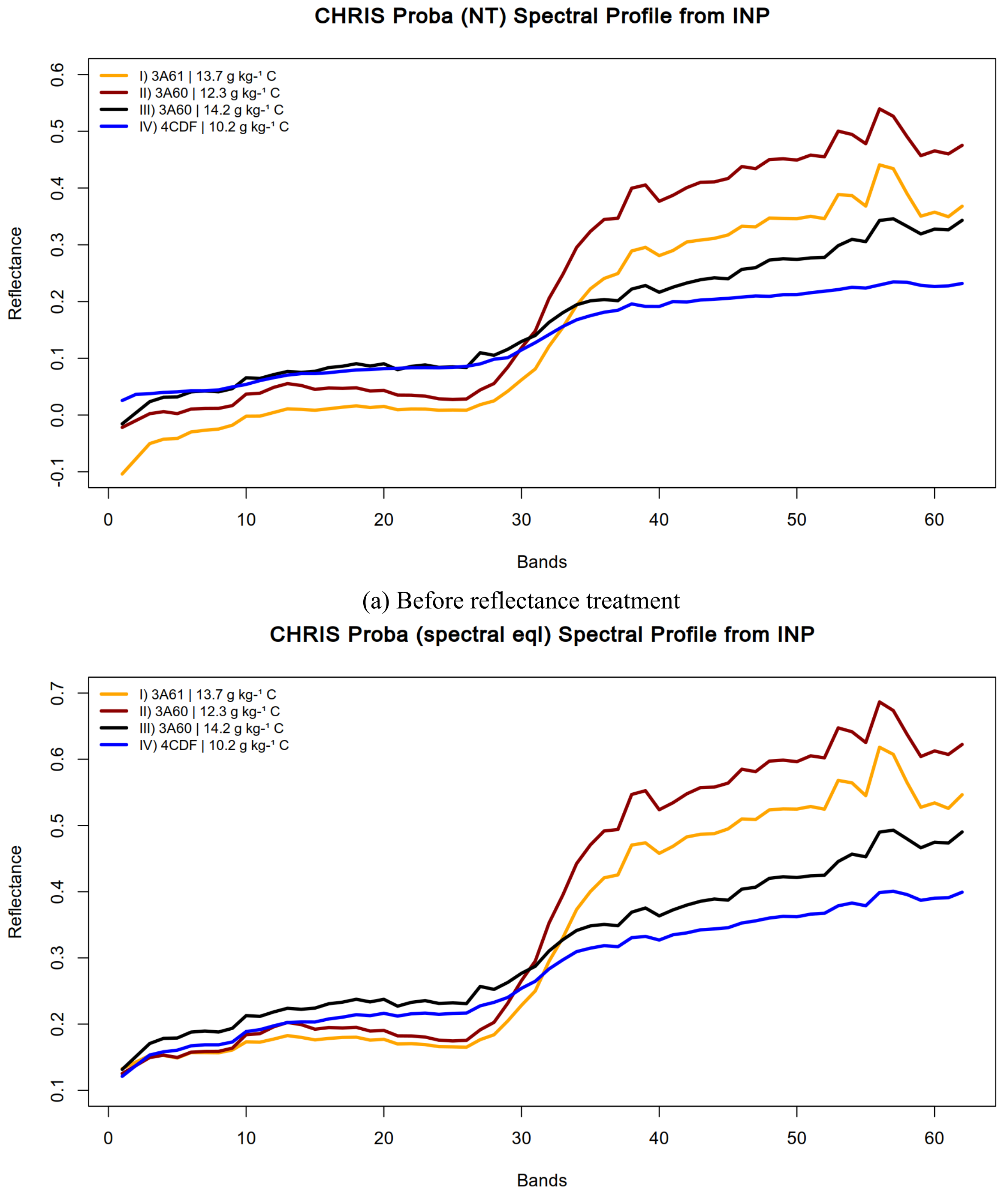
Figure 3.5: Mosaic of reflectance equalization CHRIS PROBA image of the INP plateau, Rio de Janeiro State.

and 3.5. These extracted values across the 62 bands are represented in the Figures 3.6a and 3.6b.

After the treatment, the images became visually homogeneous in terms of reflectance on the RGB plot (bands 23, 13, 3). The equalization results also reaches the NIR wavelengths, comparing Figure 3.6a to 3.6b it is possible to observe that the pixel from the southern image (4CDF) blue line had higher reflectance on the initial bands (1 to 20) with a horizontal trend. This caused a brighter image. It is also possible to highlight that the bands close to 60 (NIR) had a flat behavior, in comparison with curves of the other pixels. The Yellow line, representing the pixel from the east image (3A61), had the initial bands in the opposite situation in comparison with the previous pixel, revealing a darker image. After the treatment, all pixels had close values at band 1 (Figure 3.6b), producing a homogeneous spectral pattern (Figure 3.5). The reflectance was adjusted in a way to avoid negative values, causing a (small) shift in the ordinate axis in Figure 3.6.

3.4.4.3 Shadow analysis and adjust

The mosaicked image had large areas with strong shadows. Since the TC tends to absorb the reflectance, the shadow areas (low reflectance) tend to super estimate the amount of TC, as verified in a previous test. To decrease this influence over the TC prediction, a shadow treatment was performed. Using a shadow raster (which ranges from 0 to 1.7) derived from DEM raster, where the selecting values close to 0 deliver all pixels of the image and values close to 1.7 deliver zero pixel of shadow, the shadow values higher than 1 were selected by the expert user aiming



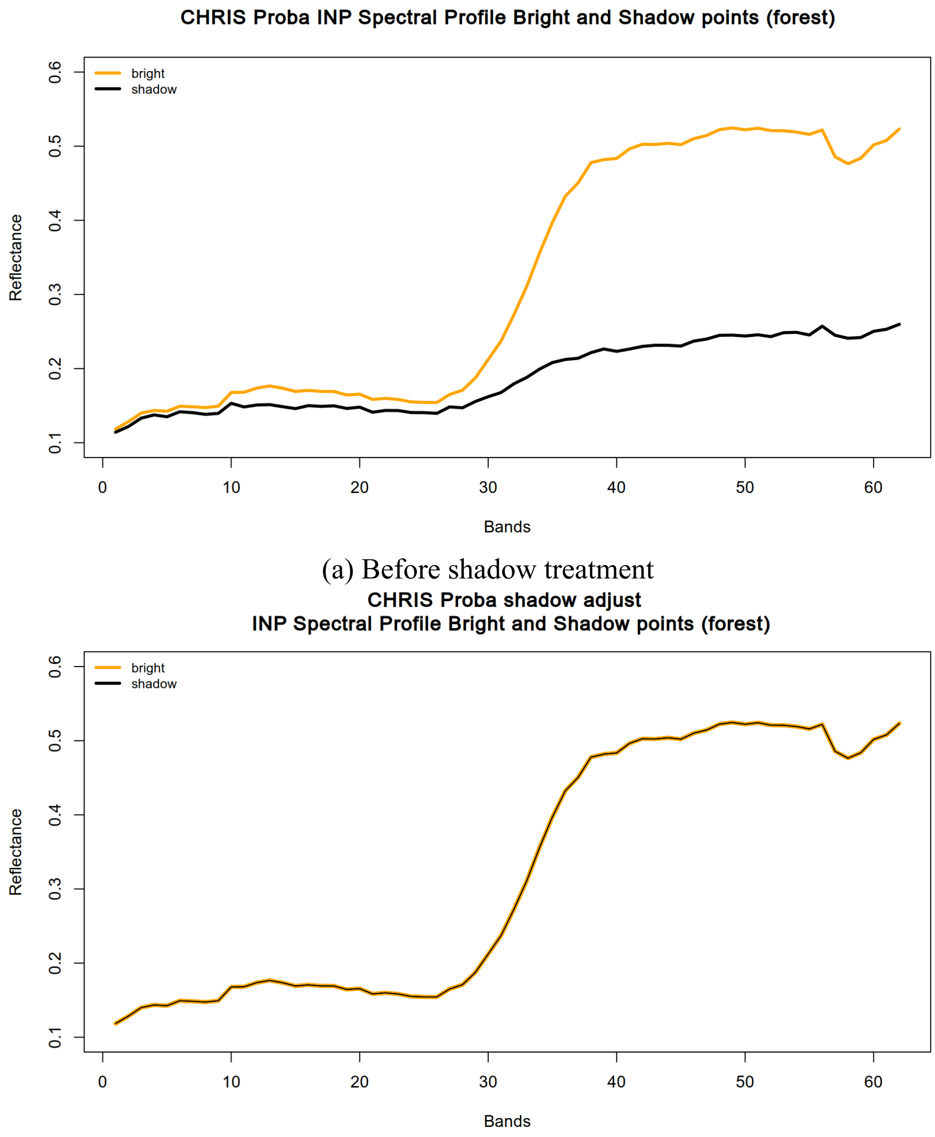
(a) Before reflectance treatment

(b) After reflectance treatment

Figure 3.6: Pixel values before (a) and after (b) the reflectance intensity equalization.

to get only pixels over the shadow areas, and the non-shadow areas were set to null value, becoming a raster mask on shadow pixels, in other words covering only the selected shadow areas and the rest of the pixels had null value. Since the DEM was obtained with 25 m surface, made by contour lines and hydrology (scale 1:50,000, IBGE data), all image data was worked to the same pixel size.

The working principle is to adjust the shadow areas of the image with its own pixel values and content, and these areas were treated with the shadow values themselves. The shadows have different intensity across the spectral bands, manly caused due the natural vegetation spectral behavior, which is higher reflectance on the NIR and lower in the Vis bands (Figure 3.7a).



(a) Before shadow treatment

(b) After shadow treatment

Figure 3.7: Spectral behavior of two pixels (control points in forest category, bright and shadow).

The INP upper part has different shadow intensity and behavior for each “island” of shadow, the first attempts of shadow adjust with a single coefficient delivered divergent reflectance over the shadows. For example, an area of shadow in forest presented reflectance of graminoid areas. Therefore, the area was subset in three categories of shadow: forest, dark forest, and graminoid areas. For each category, two control points were established, on dark and bright pixels. The spatial location of the control points pixels were chosen empirically, according to the expert knowledge. The Figure 3.7 shows the spectral behavior before and after the treatment of the control points in the forest area (category).

From band 1 to 62 specific values were incremented according to the category. The extracted shadow areas (pixels values) were compared to the bright (control point) then added to their own value times a defined β coefficient (Equation 3.1). It means each shadow pixel had its own value times an increment per band plus the original value (Equation 3.1). The intensity of the shadow treatment (β coefficient) was determined according to the expert decision over the control points per category as: β forest = *1*; β dark forest = *0.6*; β graminoid = *0.4*. The process ran in batch, as loop format for each given *j* band.

$$a_j^{band} = t_j^{band} + t_j^{band} * \beta * \frac{cpb_j^{ref} - cps_j^{ref}}{cps_j^{ref}} \tag{3.1}$$

where:

a_j^{band} is the adjusted raster band *j*

t_j^{band} is the raster band *j* to be adjusted

β is the coefficient to regulate the intensity of shadow adjust

cpb_j^{ref} is the reflectance bright control point from a given raster band *j*

cps_j^{ref} is the reflectance shadow control point from a given raster band *j*

The adjusted shadow (Figure 3.8) has recovered from very dark to close RGB values. The black spots (super estimating TC) were changed into usable values for the predictions, in consonance with Figure 3.7, where the signal was close to horizontal (shadow) and turned into bright with similar spectral behavior (vegetation).

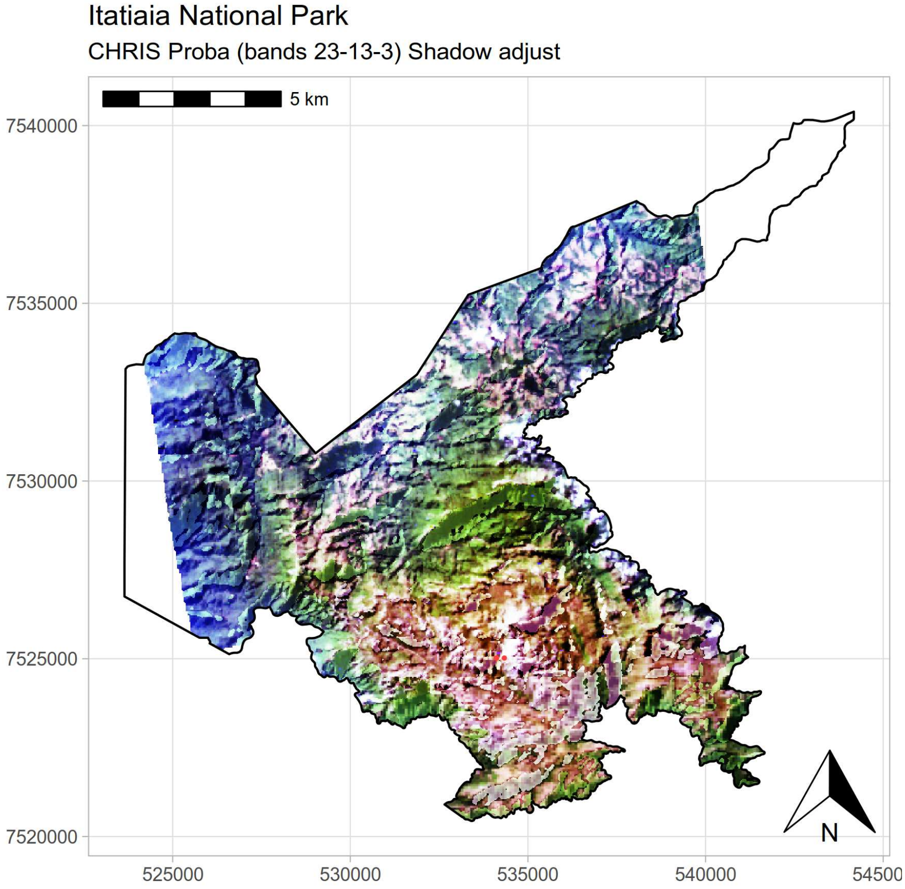


Figure 3.8: Shadow treated CHRIS image of INP plateau, Rio de Janeiro State.

3.4.5 Proximal Soil Sensing (PSS)

The Proximal Soil Sensing was conducted using 300 soil samples taken at the 84 points of INP in parallel with the study of Costa *et al.* (2020). In this Chapter we used only the top layer (top soil horizon of 84 sampled points), details about PSS are presented in the Chapter I, item 2.5.2 Soil Sampling, Analysis and Preparations. The PSS spectral range is from 350 to 2500 nm, which gives 2150 bands.

3.4.6 The Digital Soil Mapping (DSM) Process

Soil surveys require several steps, here we focus on the data processing. The process to run a Digital Soil Mapping data is characterized by application of a mathematical model and computational tools, in which, machine learning algorithms such as Random Forest learn about the environment from the behavior of the covariates. The covariates are raster data which represent environmental characteristics such as relief (and more topographic information), vegetation, geology, etc.

To illustrate the process (Figure 3.9), we can say that the model is anchored on the sampled points and the covariates are placed "below" it, in this way the model associates the values of each covariate (on the raster pixel) to each sampled point (preserving the point location). This configure the first step to train the model. Next step the model applies the "knowledge" to other pixels spread over the space of raster covariates, and based on what it was learned before it estimates a value to each pixel.

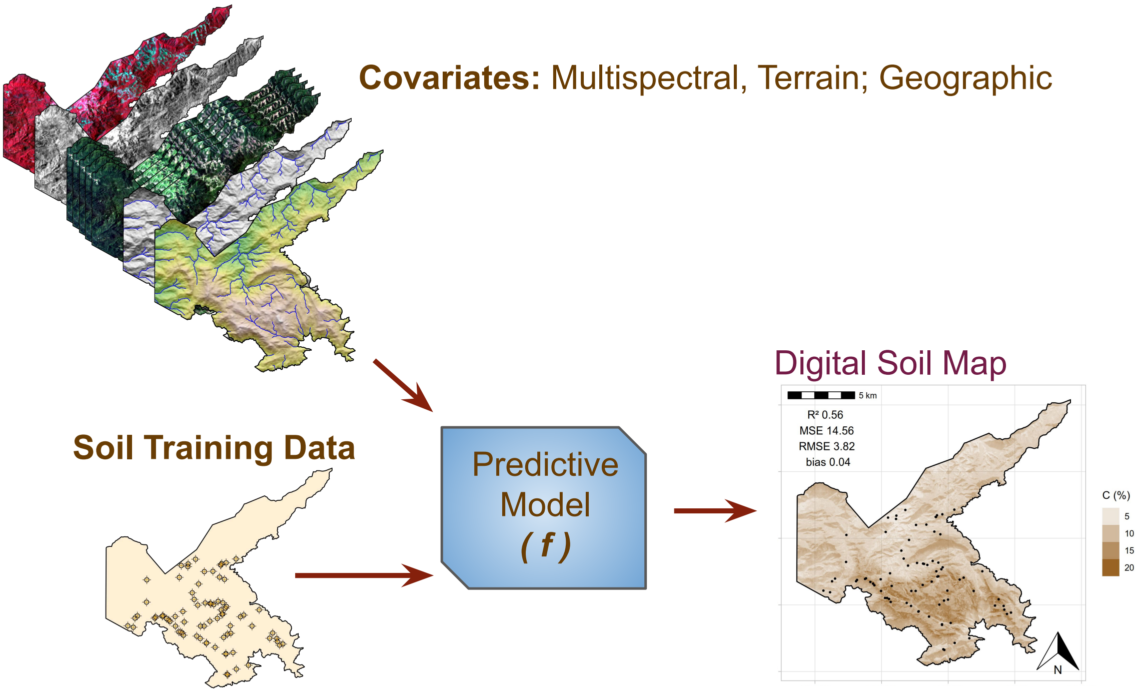


Figure 3.9: The steps of DSM processing applied to soil of INP plateau, Rio de Janeiro State.

3.4.7 Subsurfacing Image Process, the CHRIS Operation

We are calling subsurfacing an image the process to combine a Multi/Hyperspectral image with PSS using machine learning techniques. The goal is to create a new covariate set to use in DSM. In this step, the Random Forest algorithm was applied to the PSS spectrums (response variables, read in laboratory from INP soil samples), to spatially predict those spectrums. This step works similarly as in the DSM (Figure 3.9) but instead of predict a soil property, we predict a chosen wavelength spectrum (Figure 3.10). The process was repeated to each desired wavelength. As output, it was possible to have images for each wanted PSS wavelengths, giving a Subsurface Hyperspectral image. The introduction of this image into a model to map a soil property was called in this study the Hyperspectral Soil Mapping process.

As covariates (predictor variables), the final adjusted CHRIS PROBA Hyperspectral image (with 62 bands) (Figure 3.8) and RapidEye bands were used in combination with Terrain and other covariates in Table 3.1. To cover the spectral range of PSS (350 to 2500 nm), it was chosen to predict 100 wavelengths bands, in the middle of each CHRIS band (62) plus 38 bands (every 40 nm until 2500 nm) to cover the entire available PSS spectrum.

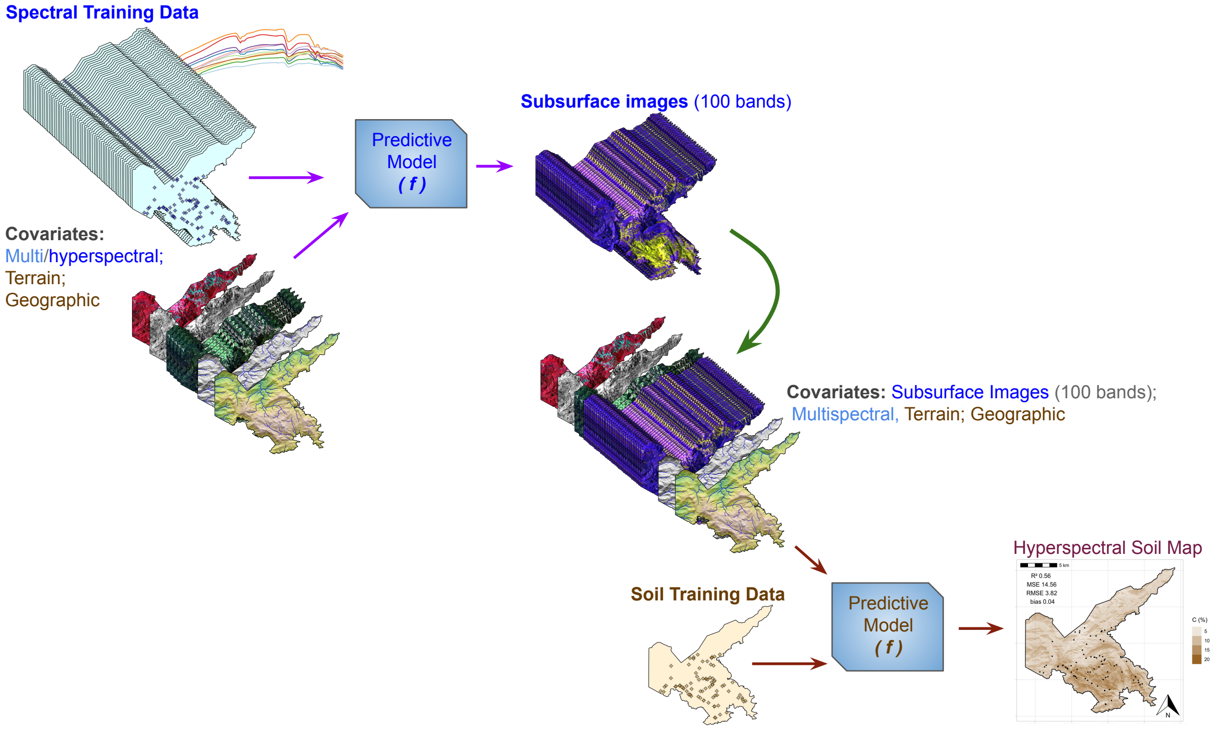


Figure 3.10: Subsurfacing process, the steps to the Hyperspectral Soil Mapping (HSM) processes applied to soils of the INP plateau, Rio de Janeiro State.

In order to match the predicted PSS wavelengths with the CHRIS bands there were used the CHRIS wavelengths values (1 to 62): 411, 442, 452, 461, 471, 481, 490, 500, 510, 520, 530, 540, 551, 561, 572, 581, 590, 603, 613, 622, 631, 641, 651, 661, 672, 680, 686, 691, 697, 703, 709, 716, 722, 728, 735, 742, 748, 755, 762, 770, 777, 785, 792, 800, 808, 833, 841, 850, 859, 868, 877, 886, 895, 905, 915, 925, 940, 955, 965, 976, 987, 997 nm. And from 63 to 100 (every 40 nm), direct from PSS: 1000, 1040, 1080, 1120, 1160, 1200, 1240, 1280, 1320, 1360, 1400, 1440, 1480, 1520, 1560, 1600, 1640, 1680, 1720, 1760, 1800, 1840, 1880, 1920, 1960, 2000, 2040, 2080, 2120, 2160, 2200, 2240, 2280, 2320, 2360, 2400, 2440, 2480 nm. Details about the spectral readings are available in the Chapter I, item 2.5.2 Soil Sampling, Analysis and Preparations.

Table 3.1: Covariates source and description.

Covariate collection	Covariate	Source	Description
Hyperspectral	CHRIS Images (Non Treated) (bands 1:62)	ESA	62 bands, initial spatial resolution of 30 meters rearranged for 25 m, spectral resolution ranging from 411 to 997 nm
	CHRIS Shadow treated (bands 1:62)	Shadow treatment of CHRIS Images	Spatial resolution of 25 m, spectral resolution ranging from 411 to 997 nm
	Subsurface CHRIS (bands 1:100)	Generated by the combination of CHRIS Shadow treatment and PSS	Spatial resolution of 25 m, spectral resolution ranging from 411 to 2480 nm
	Subsurface RapidEye (bands 1:100)	Generated by the combination of RapidEye and PSS	Spatial resolution of 25 m, spectral resolution ranging from 411 to 2480 nm
Multispectral (RapidEye)	RapidEye	RapidEye (2011)	5 bands, initial spatial resolution of 5 meters rearranged for 25 m, spectral resolution ranging from 440 to 850 nm
	NDVI		NDVI=(NIR–Red)/(NIR+Red)
	SAVI		SAVI=(1+0.5)(NIR–Red)/(NIR+Red+0.5)
Terrain	DEM	INP managers	Digital elevation model of the area-representation of the terrain's DEM INP managers 25 m surface made by contour lines and hydrology (scale 1:50,000, IBGE data)
	Slope	Derived from DEM	Gradient or rate of change of elevation between neighboring cells
	Aspect		Represents exposure faces, values in degrees (0 to 360°)
	Northernness		Indicates the direction of the slope relative to the northern. Northernness = abs(180° − Aspect)
	Plan_curv		The shape of the hillside on the horizontal plane (concave, rectilinear or convex)
	Prof_curv		The shape of the hillside on the vertical plane (concave, rectilinear or convex)
	Convergence		The general shape of the hillside in all directions (concave, rectilinear or convex)
	Cat_area		Related to volume of flooding that reaches a certain cell
	TWI		Describes a tendency for a cell to accumulate water
	LS_factor		Attribute equivalent to the topographic factor of the Revised Universal Soil Loss Equation (RUSLE)
	RSP		Represents relative slope position based on the base channel network
	CHND		Altitude above the channel network (CHNB - original elevation)
	CHNB		Interpolation of a channel network base level elevation
Geographic	Geology	(SANTOS *et al.*, 2000)	Categorical map with geological information (scale 1:50,000)
	Geomorphology	(SANTOS *et al.*, 2000)	Categorical map with geomorphological information (scale 1:50,000)

Geology classes: alluvial sediments, colluvium sediments, nepheline syenite, quartz syenite, alkaline granite, magmatic breccia, homogeneous gneisses. NDVI: normalized difference vegetation index; SAVI: soil-adjusted vegetation index; DEM: digital elevation model; Plan_curv: plan curvature; Prof_curv: profile curvature; Convergence: convergence index; Cat_area: catchment area; TWI: topographic wetness index; LS_factor: LS factor; RSP: relative slope position; CHND: channel network distance; CHNB: channel network base level. Data source: (COSTA *et al.*, 2020).

Since the spectrums are coming from laboratory measurements of samples from the upper soil horizons, taken from the first layer of the soil (not just the surface of the soil, but from a few centimeters bellow, in the middle of the horizon), they were identified as Subsurface Images. They represent more than a bare soil image, but they are a spectral subsurface image with actual spectral points anchored at sampled points. A color composition of bands 23, 13, 3 (651, 551 and 452 nm respectively) can be observed in Figure 3.12a. This RGB image of Subsurface CHRIS is in agreement with the PSS and multispectral color composite. In PSS, for example, the organic soil has lower reflectance, and the mineral soils tend to have higher reflectance (both match with the image, which tends to be dark in the areas of organic soils, and bright in the mineral soils).

The same wavelengths bands were predicted for RapidEye producing a RapidEye Subsurface image (with dozens of bands, Figure 3.12b), to compare Multispectral and Hyperspectral predictions. For both predictions terrain derivations and geographic covariates were added to the models, they are listed in Table 3.1. Similar group of covariates was used by Costa *et al.* (2020); partially used (DEM, RSP, Plan_curv, LS_factor, NDVI, slope, CHND, CHNB, geology and satellite bands) by Chagas *et al.* (2017); (DEM, geomorphology, SAVI, NDVI, slope, and satellite bands) by Pinheiro *et al.* (2019); and (DEM, Slope, Aspect, Northernness, Cat_area, TWI, NDVI, SAVI, geology and satellite bands) by Samuel-Rosa *et al.* (2015).

The four pixels from Figure 3.4 were extracted from the Subsurface CHRIS and Subsurface RapidEye. The pixels' pattern can be observed in Figure 3.13. In comparison with the spectral pattern of Figure 3.6 it turns more linear until close to the band 62, which fits in the features from soil spectrum. Also, in agreement with PSS spectral pattern (Figure 3.11, green) (with the wavelength scale proportions accommodation). The use of PSS for mapping gives a large source of information due the spectral range availability in comparison with orbital sensors (Figure 3.11).

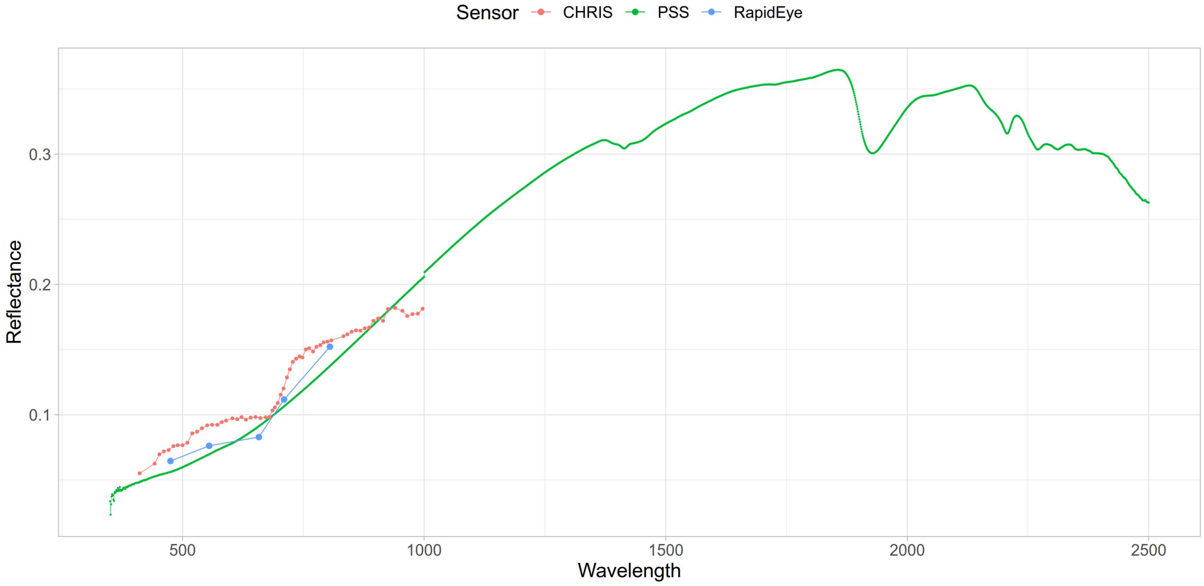


Figure 3.11: Multispectral (blue), Hyperspectral (orange) and proximal sensor (green).

3.4.8 Spatial Prediction of Total Carbon

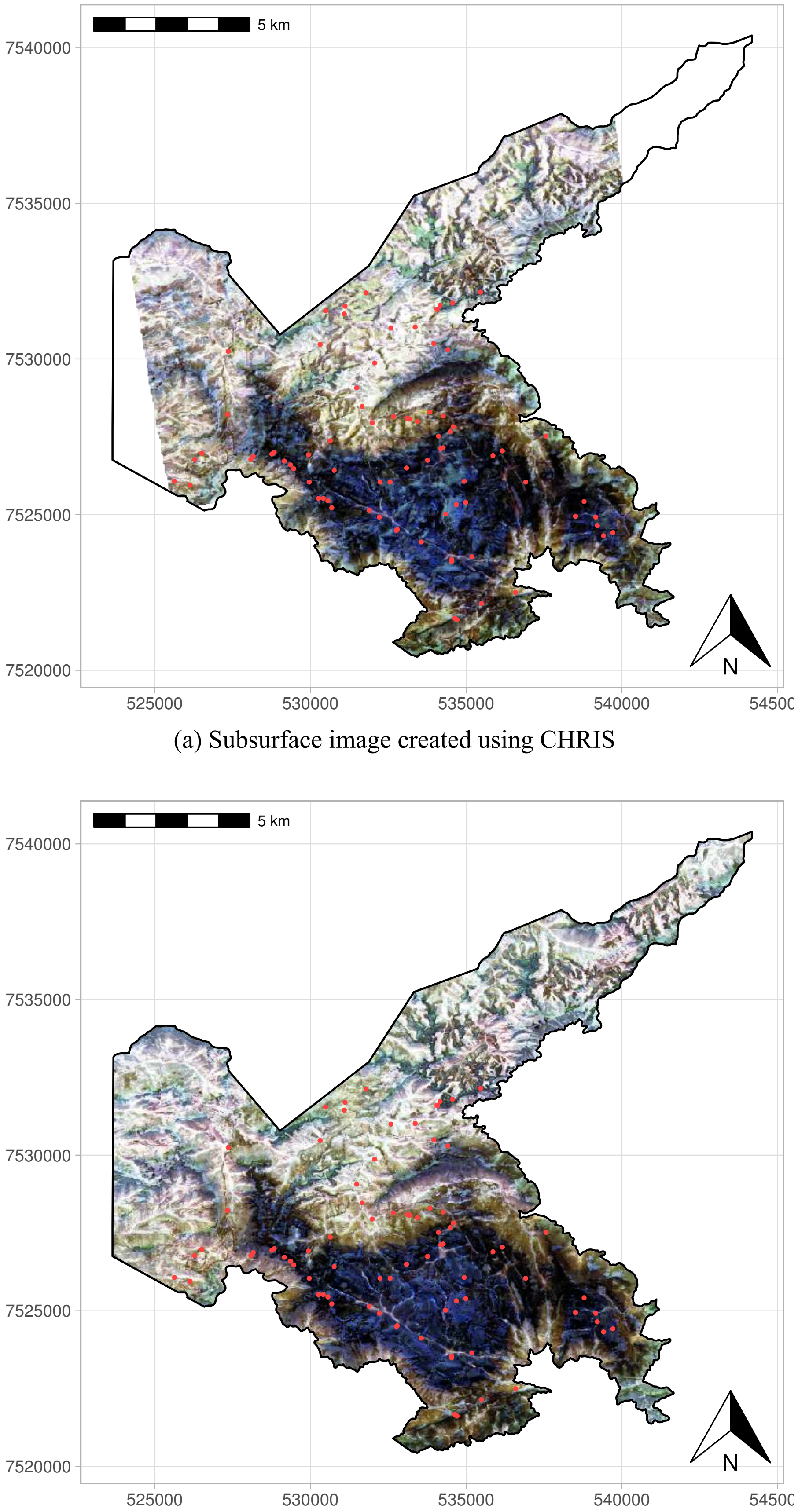
With the acquired and generated covariates (Table 3.1), Random Forest machine learning algorithm was applied to train models with the groups of covariates (Table 3.2). For each group, the spatial prediction was conducted, and to assess the results the Random Forest builtin out-of-bag (OOB) cross validation was used.

Table 3.2: Groups of covariates to predict TC on soils of INP plateau.

Covariates groups
RapidEye + Terrain + Geographic (Without CHRIS)
Non Treated CHRIS + RapidEye + Terrain + Geographic
CHRIS Shadow treated + RapidEye + Terrain + Geographic
Subsurface CHRIS (bands 1:62) + RapidEye + Terrain + Geographic
Subsurface CHRIS (bands 1:100) + RapidEye + Terrain + Geographic
Subsurface RapidEye (bands 1:100) + RapidEye + Terrain + Geographic

In summary, the process applies the 100% pure soil signal (reflectance from PSS) in soil sampled pixels from the INP, then it is used to spatial predict with the pure soil reflectance over the entire study area. With the generated pure soil reflectance image (Subsurface image), it is possible to predict the desired soil property, in this case TC. To illustrate the complete process, a workflow is presented in the Figure 3.14, plus a detailed of scheme Subsurfacing process in Figure 3.10.

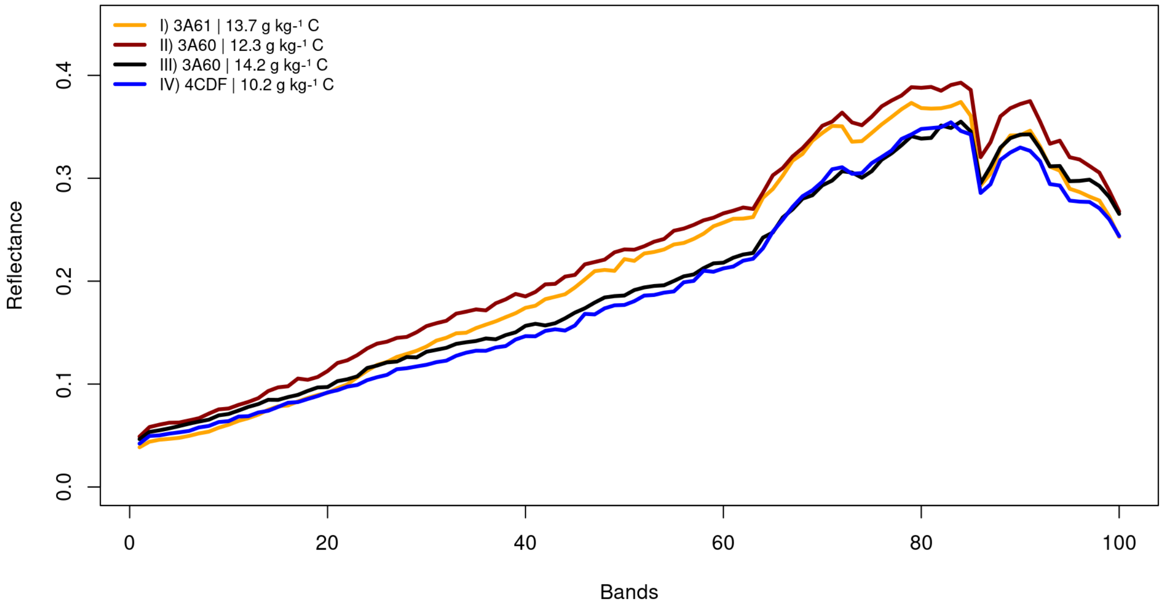
The spectral pre-processing IRF4 was applied to the Subsurface CHRIS image aiming to increase the prediction potential. Although the background tests showed little increment in comparison with the original Subsurface CHRIS image, we decided to leave the pre-treatments for the previous Chapter.



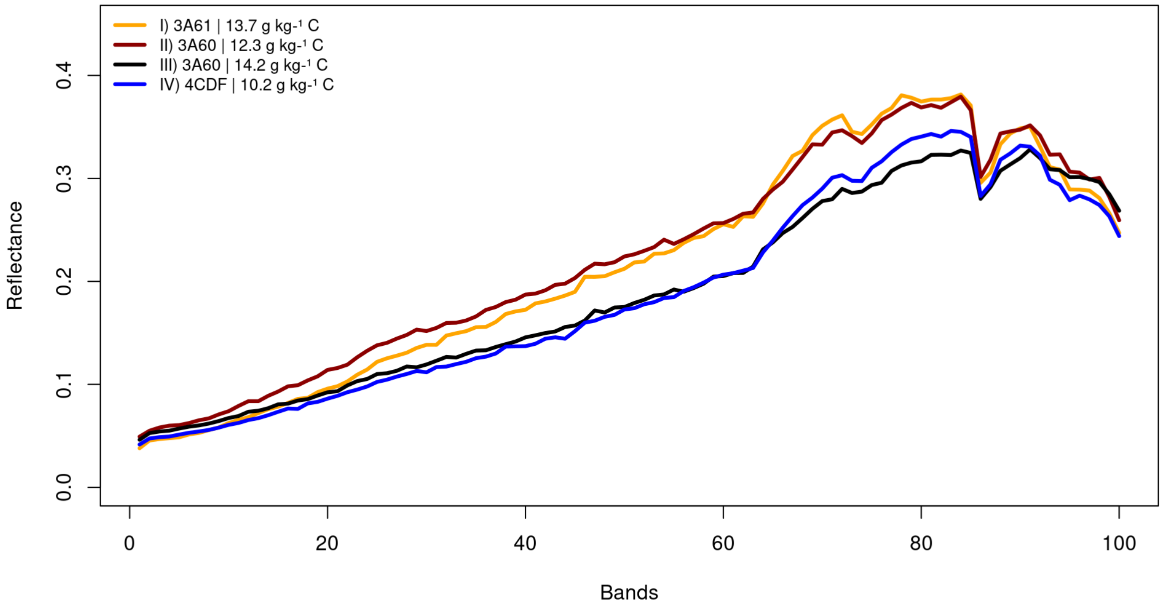
(a) Subsurface image created using CHRIS

(b) Subsurface image created using RapidEye

Figure 3.12: The upper part of INP plateau Subsurface image created using CHRIS and RapidEye (bands 23, 13, 3, corresponding to 651, 551 and 452 nm, positioned as RGB). Notice, the main difference among them is that RapidEye covers the entire area of INP plateau.



(a) Subsurface CHRIS Spectral Profile



(b) Subsurface RapidEye Spectral Profile

Figure 3.13: Comparison of spectral behavior Subsurface CHRIS and RapidEye.

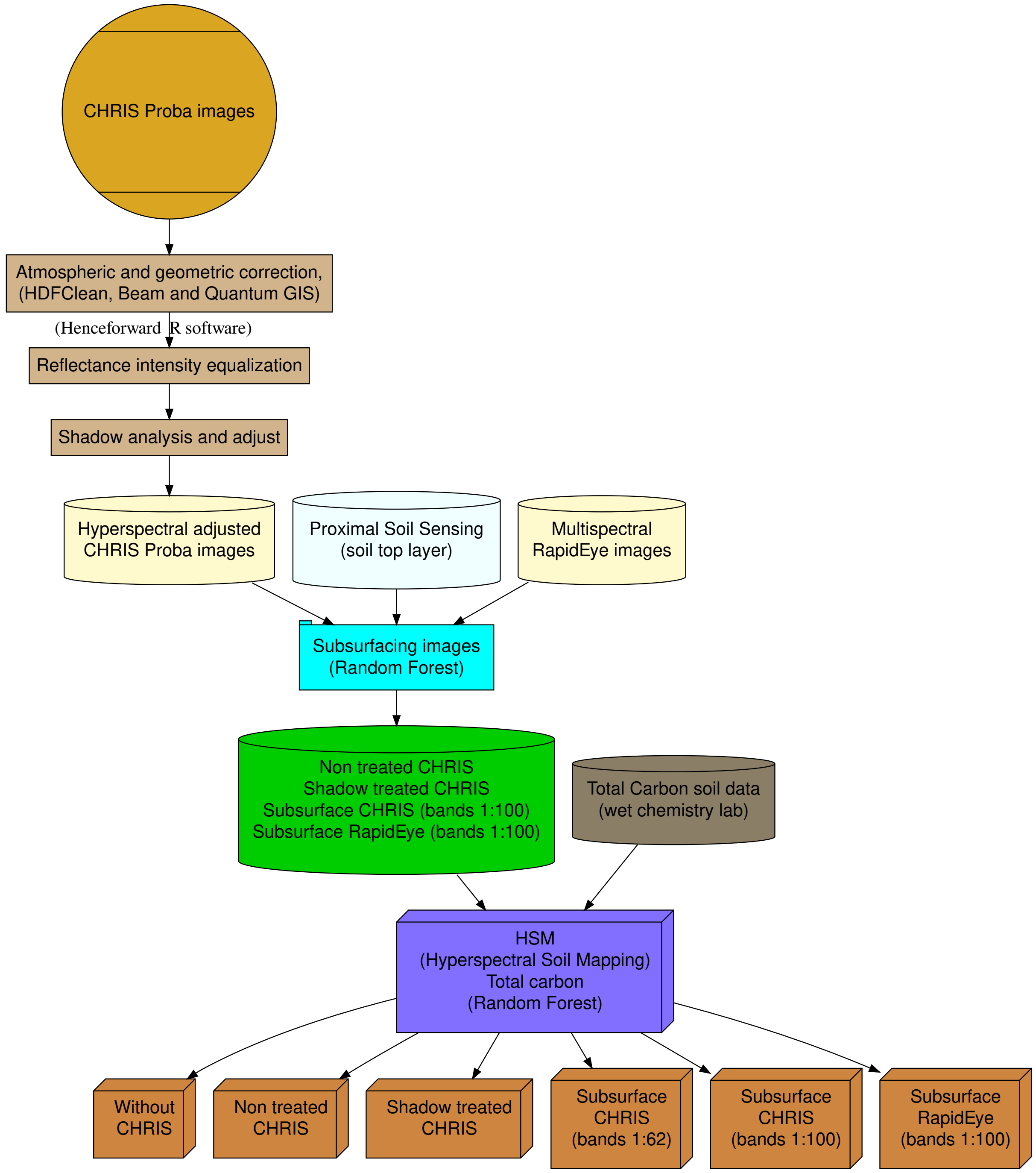


Figure 3.14: Workflow through CHRIS treatment, adjusts, modeling the Subsurface images, and spatial prediction TC with HSM in INP, Rio de Janeiro state.

3.5 RESULTS

With the application of RapidEye, Terrain, and Geographic covariates (without CHRIS) the spatial prediction of TC produced the following descriptive statistics indexes: coefficient of determination (R^2) 0.32, Mean Squared Error (MSE) 22.39%, Root Mean Square Error (RMSE) 4.73% and bias 0.19 (Figures 3.15a and 3.15b). The 5 higher ranked covariates by the model were DEM, Northness, CHNB, RapidEye bands 5 (Near IR) and RSP (Figure 3.15c).

Using the Non Treated CHRIS image, the predicted TC plot (Figure 3.16a) presented different amounts of TC for each image (3A60, 3A61, 4CDF), resulting in a segmented map, with a strong relationship with the Non Treated CHRIS (Figure 3.4). The descriptive statistics shows values of R^2 0.33, MSE 21.9%, RMSE 4.68% and bias 0.2 (Figures 3.16a and 3.16b). The top of ranked covariates are dominated by the groups Terrain (DEM 1st), Spectral (RapidEye) and Geographic. The Hyperspectral covariates start to appear in the 8th position (Figure 3.16c).

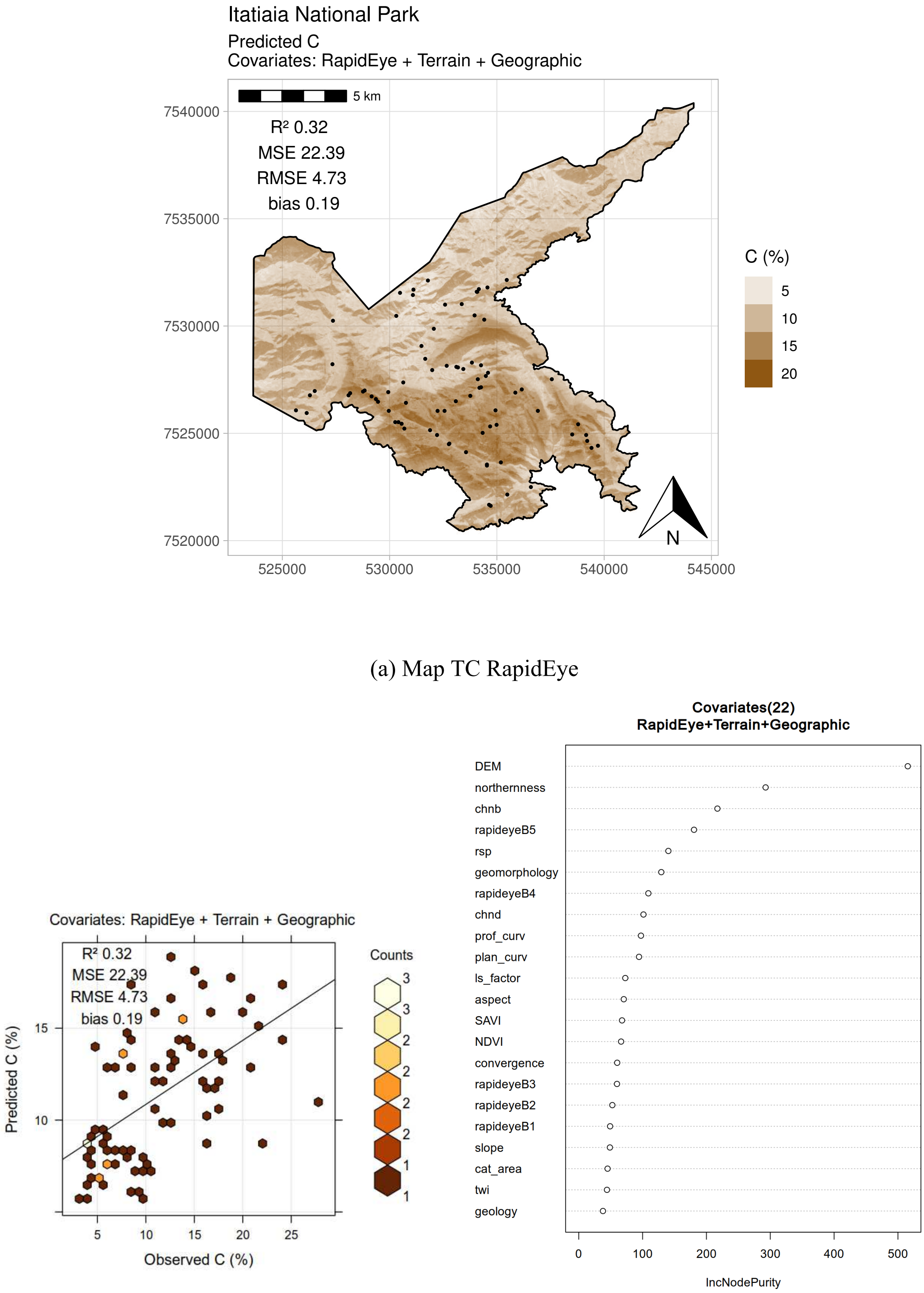
The CHRIS image with the shadow treatment resulted in a homogeneous plot map and higher statistical coefficient in comparison with Non Treated CHRIS; with R^2 0.36, MSE 20.96%, RMSE 4.58% and bias 0.24 (Figures 3.17a and 3.17b). The predominant group of covariates in the ranked list is similar to Non Treated CHRIS, starting with DEM as well. The Hyperspectral covariate starting in the 6th position (Figure 3.17c).

The Subsurface CHRIS (using layers 1 to 62) is a changing perspective, with the descriptive statistics outstanding as R^2 0.58, MSE 13.8%, RMSE 3.71% and bias 0.12 (Figures 3.18a and 3.18b). The covariates in the ranked list are mostly from Hyperspectral group (subsurface CHRIS), starting with band 53. The Relief appears in 6th, 19th and 22nd position with Northness, Prof_curv and DEM respectively (Figure 3.18c).

In sequence, Subsurface CHRIS using the plus 38 predicted wavelength (using layers 1 to 100), had the descriptive statistics value of R^2 0.57, MSE 14.13%, RMSE 3.76% and bias 0.09 (Figures 3.19a and 3.19b). Among the first 40 covariates only three covariates are from other groups else then Hyperspectral. Northness is in 2nd position, RapidEye band 4 in 32nd and plan_curv in 39th. The subsurface spectral bands (with the wavelengths, from 1000 to 2480 nm, above the nominal values of the original CHRIS PROBA, bands 63 to 100) were ranked in the 7th to the 9th position, bands 75 and 71 respectively (Figure 3.19c).

The Subsurface RapidEye (using layers 1 to 100), had the descriptive statistics value of R^2 0.56, MSE 14.56%, RMSE 3.82% and bias 0.04 (Figures 3.20a and 3.20b). The first ranked covariate is the subsurface spectral band 75, with Northness as 3rd one. Among the first 40 covariates, again, only three covariates are from other groups than Hyperspectral (Figure 3.20c). From figures 3.15b to 3.20b, the transversal line is the fitted model line.

The summary of the spatial prediction of TC at the Itatiaia National Park upper plateau is presented in Table 3.3. The R code and data can be found on the repositories of the Author.

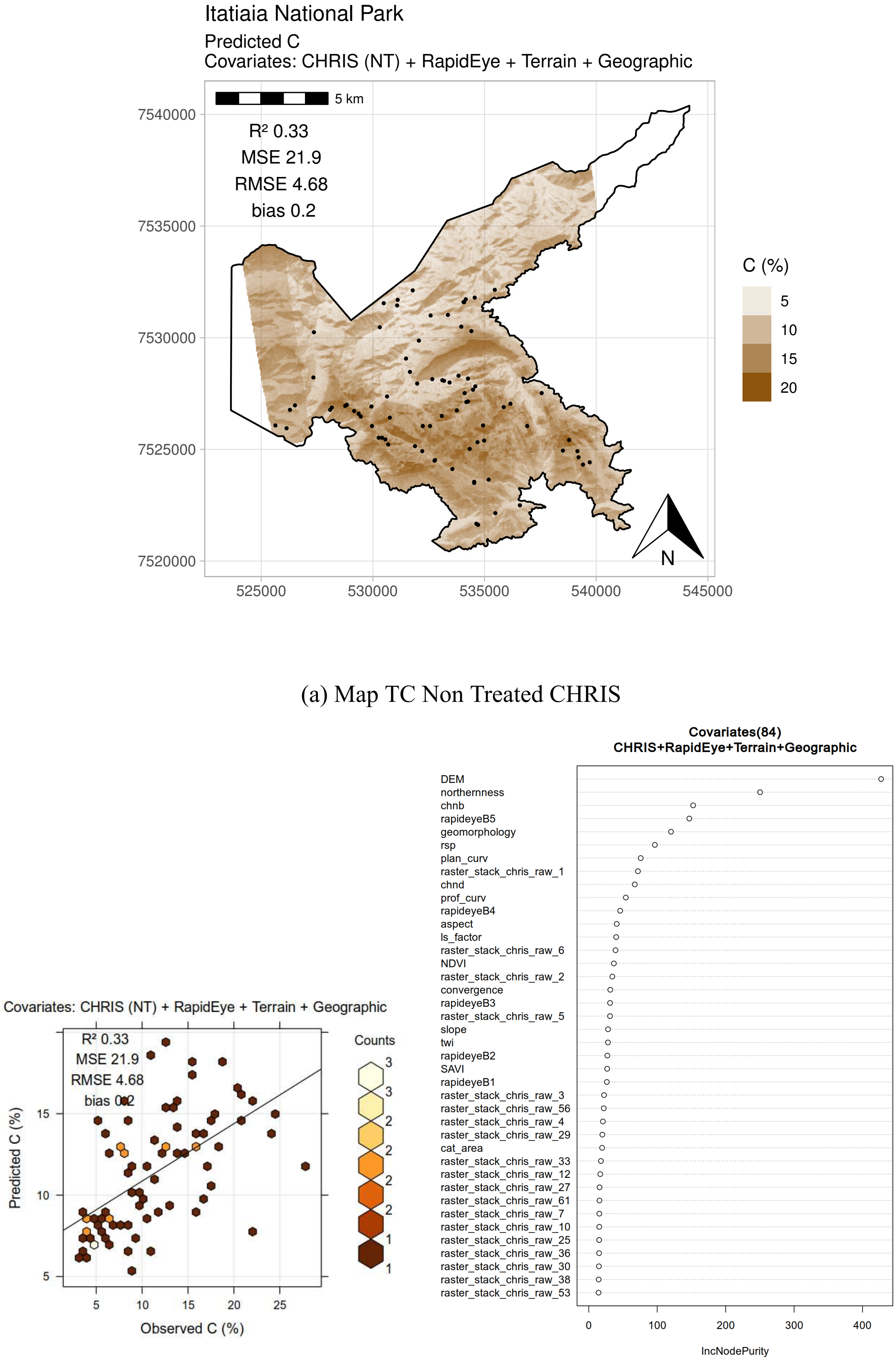


(a) Map TC RapidEye

(b) TC RapidEye, observed versus predicted

(c) Model covariates ranking

Figure 3.15: Spatial prediction of TC over INP plateau, with the covariates RapidEye, Terrain, Geographic.

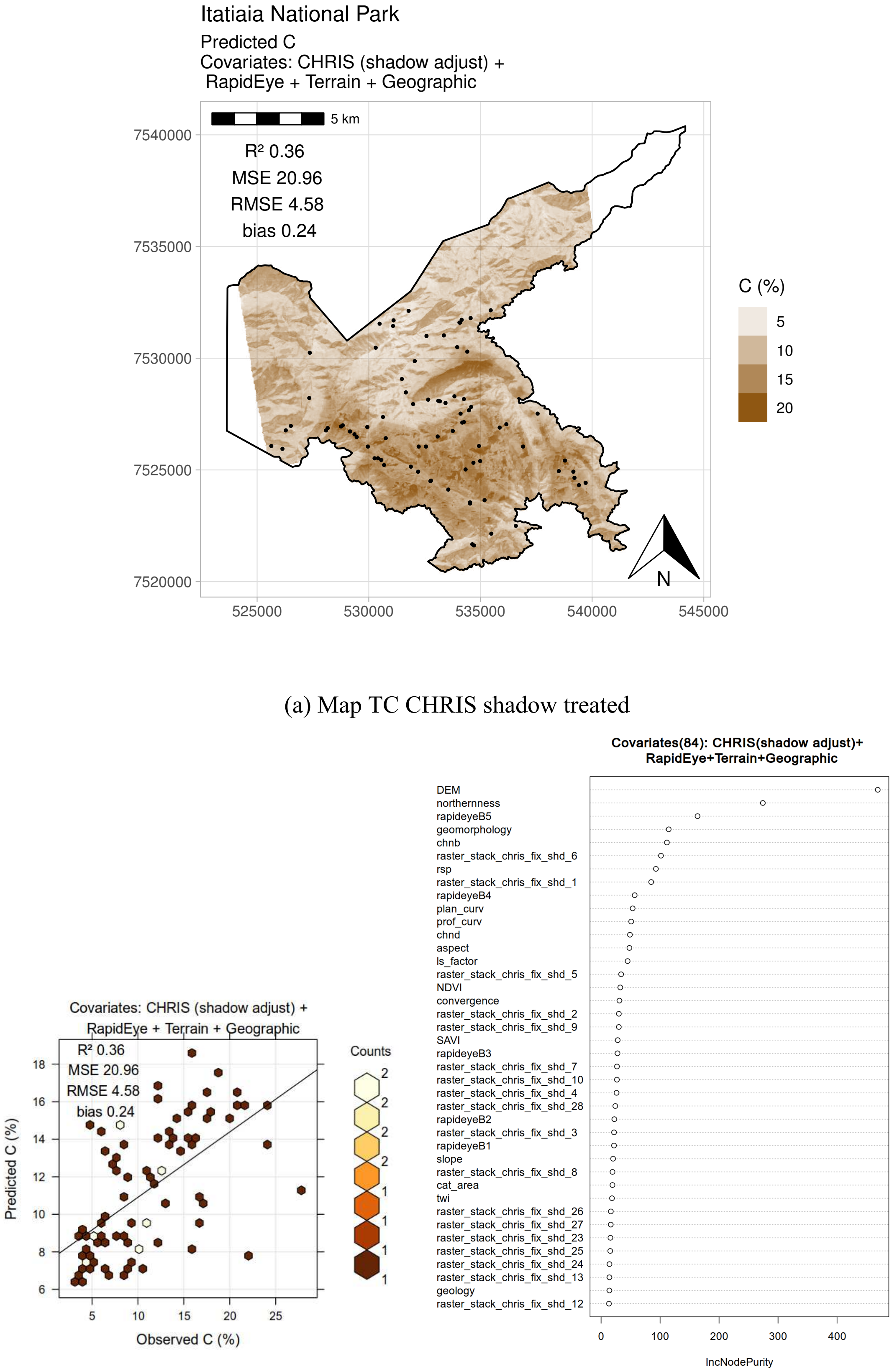


(a) Map TC Non Treated CHRIS

(b) TC Non Treated CHRIS, observed versus predicted

(c) Model covariates ranking

Figure 3.16: Spatial prediction of TC of soils in the INP plateau, with the covariates Non Treated CHRIS, RapidEye, Terrain, Geographic.

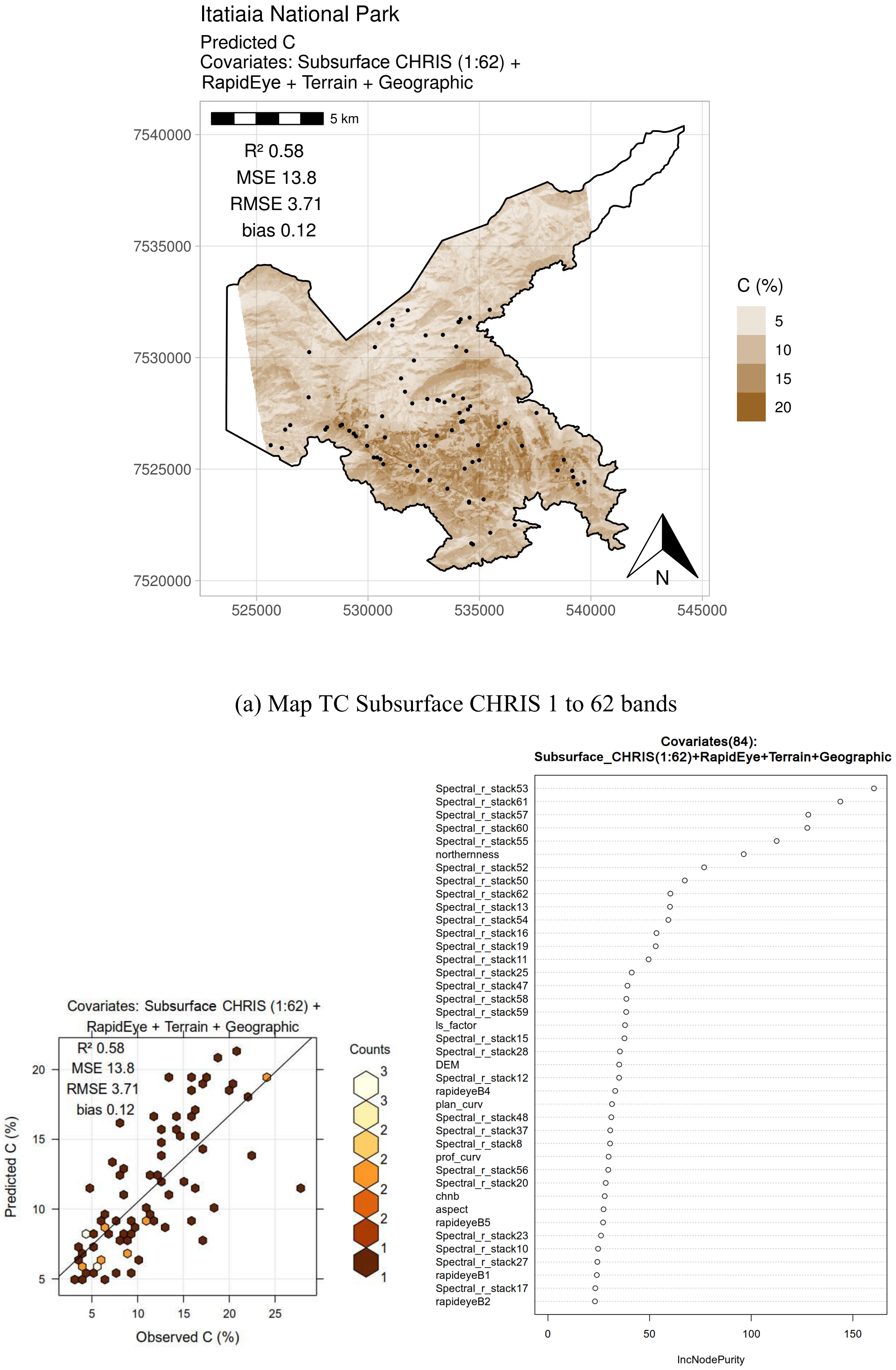


(a) Map TC CHRIS shadow treated

(b) TC CHRIS shadow treated, observed versus predicted

(c) Model covariates ranking

Figure 3.17: Spatial prediction of TC of soils in the INP plateau, with the covariates Shadow treated CHRIS, RapidEye, Terrain, Geographic.

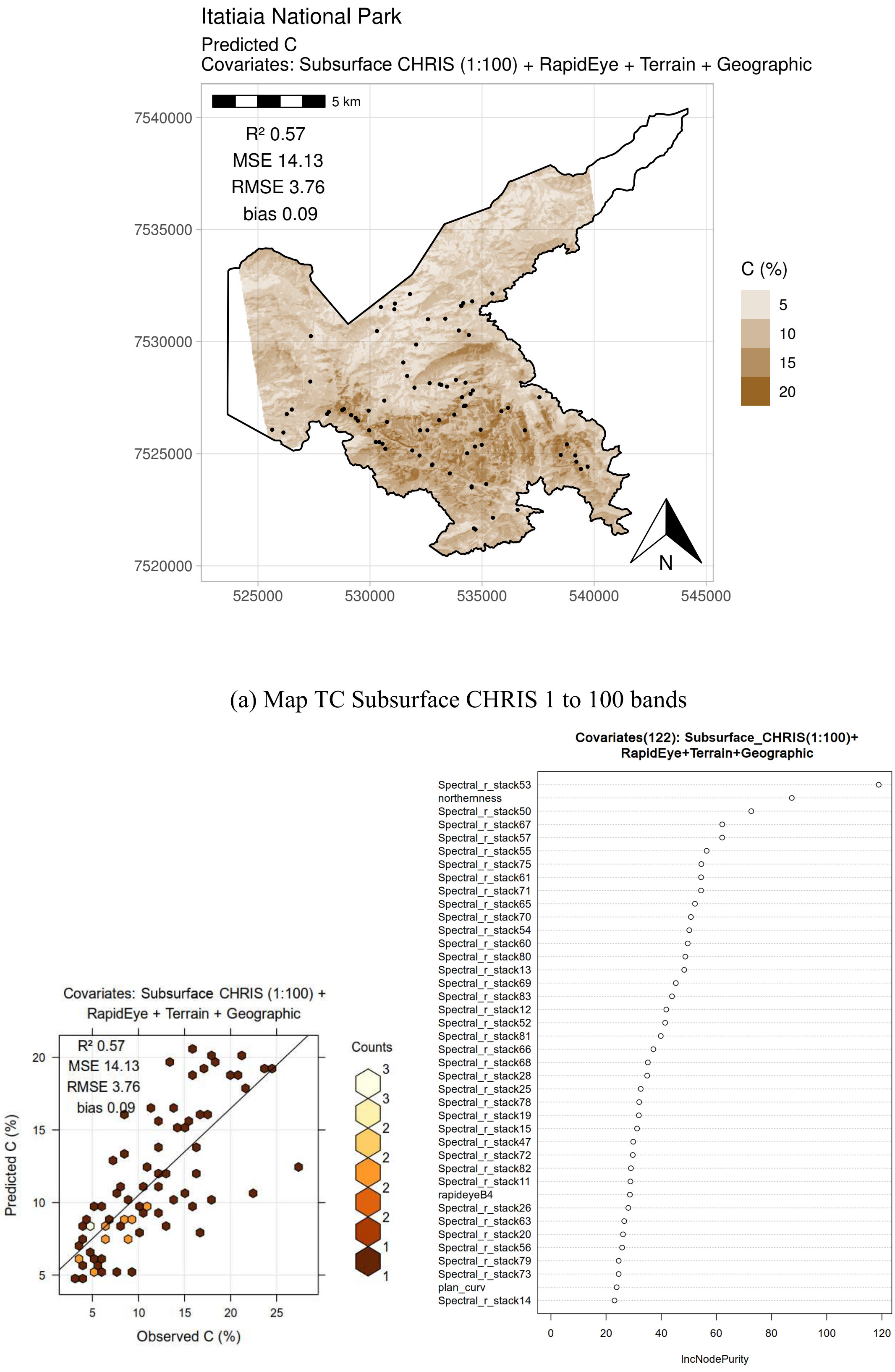


(a) Map TC Subsurface CHRIS 1 to 62 bands

(b) TC Subsurface CHRIS 1 to 62 bands, observed versus predicted

(c) Model covariates ranking

Figure 3.18: Spatial prediction of TC of soils in the INP plateau, with the covariates Subsurface CHRIS (bands 1 to 62), RapidEye, Terrain, Geographic.

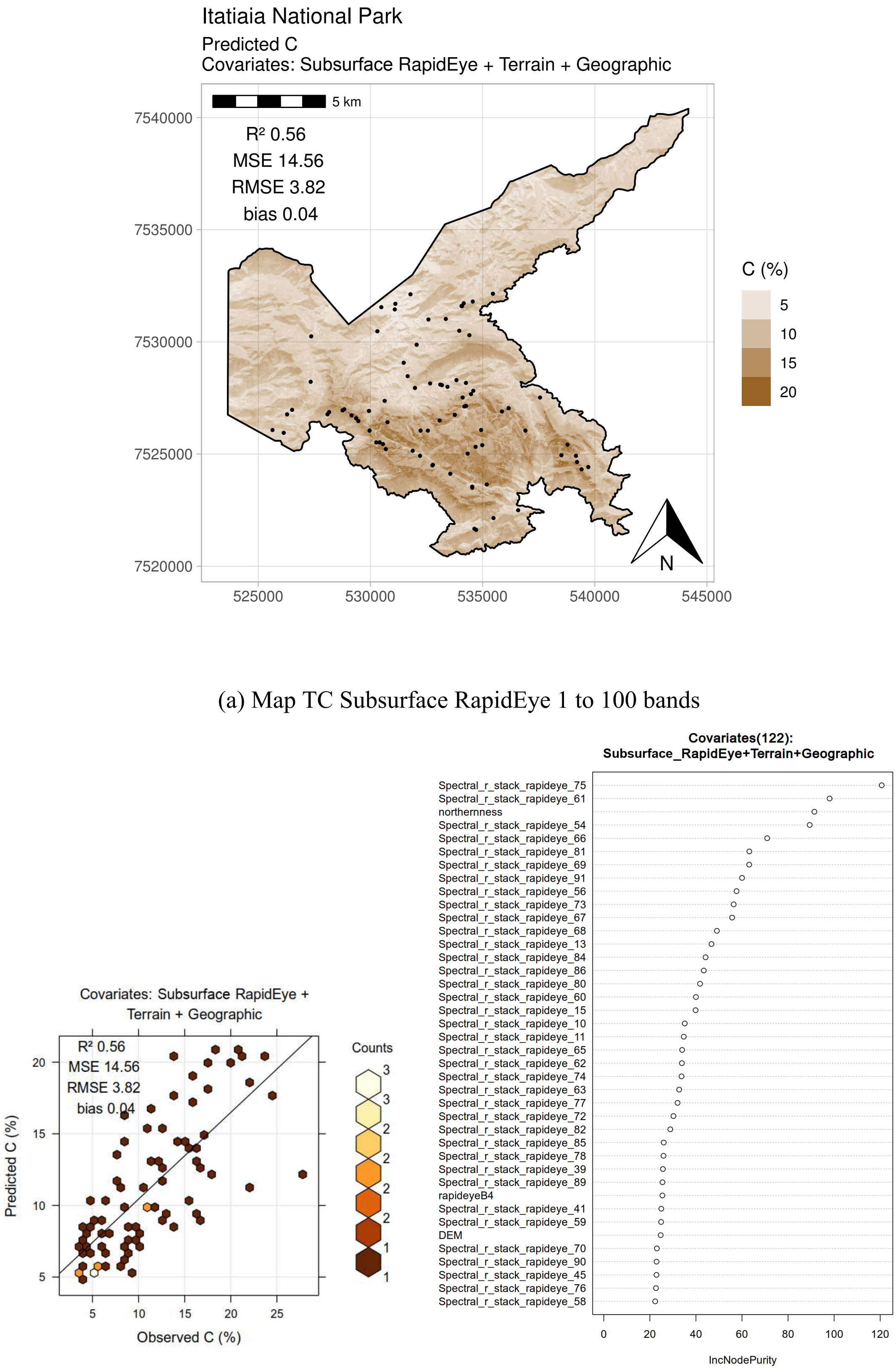


(a) Map TC Subsurface CHRIS 1 to 100 bands

(b) TC Subsurface CHRIS 1 to 100, observed versus predicted

(c) Model covariates ranking

Figure 3.19: Spatial prediction of TC of soils in the INP plateau, with the covariates Subsurface CHRIS (bands 1 to 100), RapidEye, Terrain, Geographic.



(a) Map TC Subsurface RapidEye 1 to 100 bands

(b) TC Subsurface RapidEye 1 to 100 bands, observed versus predicted

(c) Model covariates ranking

Figure 3.20: Spatial prediction of TC of soils in the INP plateau, with the covariates Subsurface RapidEye (bands 1 to 100), RapidEye (multispectral bands 1 to 5), Terrain, Geographic.

Table 3.3: Descriptive statistics of spatial prediction of TC of soils in the INP plateau, covariates used, and the main explanatory covariate.

Covariates groups	R^2	MSE	RMSE	bias	Main cov.
RapidEye + Terrain + Geographic (Without CHRIS)	0.32	22.39	4.73	0.19	DEM
Non Treated CHRIS + RapidEye + Terrain + Geographic	0.33	21.9	4.68	0.2	DEM
CHRIS Shadow treated + RapidEye + Terrain + Geographic	0.36	20.96	4.58	0.24	DEM
Subsurface CHRIS (bands 1:62) + RapidEye + Terrain + Geographic	0.58	13.8	3.71	0.12	band 53
Subsurface CHRIS (bands 1:100) + RapidEye + Terrain + Geographic	0.57	14.13	3.76	0.09	band 53
Subsurface RapidEye (bands 1:100)+ RapidEye+ Terrain+ Geographic	0.56	14.56	3.82	0.04	band 75

R^2: coefficient of determination; MSE: Mean Squared Error in (%) as TC; RMSE: Root Mean Square Error in (%) as TC; Main cov.: Higher ranked covariate from correspondent Random Forest model. Band 53 = 895 nm and band 75 = 1480 nm.

3.6 DISCUSSION

Considering that the semivariogram presented a spatial dependence of 750 m for TC sampling of soil in the INP plateau, it is acceptable that 84 soil profiles were sufficient to perform the spatial modeling. A study of soil carbon in an area located in the South region of Brazil by Samuel-Rosa *et al.* (2015) used a much more populated mesh points of 350 samples in $\sim$2000 hectares, when compared to INP plateau (84 points for 16402 hectares), and the semivariograms have similar aspects which confirm that the 84 points are sufficient for this study area.

Comparing the prediction without CHRIS (covariates: RapidEye, Terrain, Geographic) and the Non Treated CHRIS, the second had slightly better coefficients (probable due to usage of CHRIS bands in the model), although, as expected, the map plot still reflected the mosaic differences originated from the source image (Non Treated CHRIS). The mosaic has darker area over the image 3A61 (lower reflectance) and lighter on 4CDF (higher reflectance) in comparison with the 3A60. This reflectance differences were captured by the Random Forest model, as presented in the TC map in Figure 3.15a.

The shadow treatment is a process that uses the raw data from each pixel, as a result it tends to reduce the super estimation of TC in the shadow areas. This Hyperspectral image presented improvements in comparison with predictions without CHRIS and Non Treated CHRIS. The R^2 reached 0.36 in comparison with 0.32 and 0.33 from previous prediction. A visual inspection of the map plots indicated lesser super estimation over TC amounts for the shadow treatment. For the three first prediction the DEM covariate is at the top of the ranking, and the Terrain group is usually higher in the rank. Analogous to Chagas *et al.* (2017), where elevation (DEM) was highly ranked with Random Forest, to predict soil classes, and to map phytophysiognomies in Pinheiro *et al.* (2019). Mapping TC in a Chinese province, with boosted regression trees algorithm, Wang *et al.* (2017) had the DEM as most important covariate. This covariate builds a strong relation with TC, in accordance with this study, unless when used the Subsurface Image as covariate.

It is noticeable that the prediction with Multispectral image SAVI and NDVI ranked together (Figure 3.15c). However, with Non Treated CHRIS and shadow treatment, NDVI was higher ranked, eight and four position, respectively (Figures 3.16c and 3.17c). Regarding to the Subsurface images (in HSM) both were placed after the fortieth position (out of the plotted figure range, Figures 3.18c, 3.19c and 3.20c).

The sampled points were mainly in the bright areas (see Figure 3.5), leading to minimize the effects of the shadow treatment, when translated into validation coefficients. The treatment can be improved by picking more values from different shadow points across the entire image and treating each shadow (or part of it) individually.

The Subsurface CHRIS (1 to 62 and 1 to 100) resulted in a map with lighter color tones, also in the shadow areas lesser dark colors are observed, which indicates better estimation of TC in comparison with the previous plots. The TC prediction which used Subsurface covariates had much larger presence of Hyperspectral covariates (subsurface covariates) and higher ranked among the most useful in the models than the others predictions. The difference between the application of wavelengths, from 411 to 1000 nm (62 bands) or 411 to 2500 nm (100 bands), was not significant in this study. However, it is worth to test across the V-SWIR wavelength range, since it might behave better with a larger sample size.

TC prediction with Subsurface RapidEye (with 100 bands) was above the expectations, getting close to the Subsurfacing CHRIS. In this case, the most important covariate was the band 75 (1480 nm), and in the rank of the important covariates, 22 from 40 are from wavelength higher than 997 nm (last band of CHRIS in NIR, edge of wavelengths), and 30 of 40 above the

NIR limit of RapidEye 850 nm. The ranked covariates (Figure 3.20c) from band 75 (on top) with the correspondent wavelengths are: 1480, 987, northernness, 905, 1120, 1720, 1240, 2120, 925, 1400, 1160, 1200, 551, 1840, 1920, 1680, 976, 572, 520, 530, 1080, 997, 1440, 1000, 1560, 1360, 1760, 1880, 1600, 762, 2040, RapidEye Band 4 (760-850), 777, 965, DEM, 1280, 2080, 808, 1520 and 955 nm. It shows the extraordinary gain of the prediction using the entire available range of wavelengths. The model ranking of the covariates showed the advantage of exploring the covariates in further wavelengths, such as those provided by the PSS used in this study, and it is reported that the MID infrared wavelengths, in PSS have better potential to predict soil properties than V-SWIR (ROSSEL *et al.*, 2005; DANGAL *et al.*, 2019). Thus, we presume that application of MID infrared might improve even more the capacity of soil properties spatial prediction with the HSM method.

The study area in the INP plateau was the object of soil digital mapping (COSTA *et al.*, 2020), and the authors used few covariates from the *covariate selection approach* to enhance the prediction process of TC, obtaining a R^2 of 0.45. In this baseline with Multispectral covariates this study obtained the R^2 as 0.32, showing that it could be improved. Then, by using the Subsurface image, it was possible to reach the R^2 of 0.58, which confirms the potential of the HSM method. Thus, the covariate selection method still might improve the use of Subsurface image.

A mapping of an Australian region (GOMEZ *et al.*, 2008) shows that the spectral resolution is important, but it is not the major factor to obtain an accurate prediction. This was also verified in this study, when the application of 62 and 100 bands of the Subsurface image resulted in values of R^2 as 0.58 and 0.57, respectively. Similar perspective was observed in the results of previous Chapter, when it was used 299 instead of 2150 spectral covariates.

A study mapping Soil Organic Carbon (SOC) in Illinois (JABER *et al.*, 2011) found that in comparison with Multispectral, the Hyperion Hyperspectral image marginally enhanced the prediction of SOC. In a comparison of Multispecrtal and Hyperspectral images, Castaldi *et al.* (2016) found that Hyperspectral also improved TC and texture predictions in soils. In this study, a similar situation occurred with the use of Hyperspectral sensor CHRIS, and RapidEye increased the model prediction capacity from R^2 of 0.32 (RapidEye) to 0.36 (CHRIS after the reflectance equalization Figure 3.5 and shadow treatment Figure 3.8). Hyperspectral images showed better capacity to identify saline soils than the Multispecrtal images (MOREIRA *et al.*, 2015; NETO *et al.*, 2017). In this study, just to emphasize, the Subsurface images had much better results than the “simpler” Hyperspectral images.

In agreement with Jaber *et al.* (2011), the higher spatial resolution could lead to an improvement of the technique, since it should reduce the effect of mixed pixel. The effort to apply more detailed covariates should be balanced with more field sampling and analyzed for each case. According to Samuel-Rosa *et al.* (2015) for some soil properties the difference is marginal, while for TC, it may be worth. In agreement with Guevara *et al.* (2018), the expert opinion is needed to balance and check the models, avoiding unrealistically high modeling estimates of soil properties such as TC. The model tends to super estimate TC depending on the chosen covariates.

As pointed by some authors, the selection of covariates with their advantages can further improve the model’s prediction capacity (JABER *et al.*, 2011; NUSSBAUM *et al.*, 2018; GOMES *et al.*, 2019; COSTA *et al.*, 2020; WADOUX *et al.*, 2020). In this study, we focused into developing the Hyperspectral Soil Mapping method and to compare with the conventional DSM.

Studies with airborne sensors, such as Ben-Dor *et al.* (2002), Selige *et al.* (2006), Stevens *et al.* (2006), Guo *et al.* (2019), showed good spatial prediction of soil carbon, with R^2 values as 0.83, 0.90, 0.85 and 0.54, respectively, to the cited authors. The average altitude of airplane missions is of three kilometers, while satellites are in the range of hundreds of kilometers, varying from 552 to 685 km for PROBA satellite. So, airborne sensors tend to have less distortion, noise and interference, such as the atmospheric ones which are a driver for better results. Sample size and other site characteristics play a role in DSM; in this sense coefficients comparison may not reflect the model potential. In this study, we focus in the development of the HSM method with its leap improvement.

The synthetic temporal image series from Landsat 5 of Demattê *et al.* (2018), Gallo *et al.* (2018), Mendes *et al.* (2019), Padilha *et al.* (2020), the spectral-temporal response surface Zhang *et al.* (2017), Lamichhane *et al.* (2019), and similarly the bare ground images mentioned by Wadoux *et al.* (2020), all work on the level of bare soil reflectance. They are a great achievement, but we understand that the HSM (on Subsurface Image) is more efficient due to the pure PSS signal, and the possibility to expand of orbital/aerial bands to the limits of PSS wavelengths. As covariate to Subsurface the synthetic images might increase performance of the HSM.

Among the many characteristics that can affect the spatial prediction of soil properties we highlight the number of samples to validate the models, which is even harder to take in mountain regions with limited access and locomotion such as the Itatiaia National Park. Besides, in the INP the relief may influence the optical images, causing shadows and different degrees of reflectance, and it affects vegetation coverage and associated indexes, resulting in different degrees of homogeneity of the surface.

The process to combine PSS and DSM (Subsurfacing) can generate dozens, hundreds or thousands of spectral bands (as the user choose from the pool of available PSS wavelengths) to be used as new covariates with HSM, aiming toward the most accurate soil map.

3.7 CONCLUSIONS

The combination of Multi–Hyperspectral images and spectral data and PSS proved to be an improved mapping method, named here as HSM, since it improved the results of TC prediction in 75%, rising the R^2 from 0.33 to 0.58 in the case of Non Treated CHRIS to Subsurface CHRIS. For the Subsurface RapidEye it also increased in 75% the R^2 from 0.32 to 0.56. The Subsurfacing process made the images easier to be worked by the model associate with TC, we presume that it is a result of the reduction or elimination of atmospheric, land cover, vegetation interferences, and pure PSS signal in the system. The atmospheric corrections can reduce noise effects on the image and PSS had no effect of atmosphere. Thus the PSS samples are from few centimeters below ground, instead of a top soil surface, thus, the Subsurfacing process deliver literally a subsurface image with zero or almost zero atmospheric disturbance (almost zero comes from the use of orbital images, which even treated may carry residual atmospheric disturbance). Even in bare soils the Subsurfacing process is superior to any satellite or synthetic image, because it allows to expand the bands (beyond) of any optical orbital sensor (satellite) to any desire PSS band.

The HSM method presented in this study is the first direct integration between PSS and DSM, and such major improvement is rarely found in DSM techniques. Thus, following the line of chapter I, the same may occur with other soil properties and should be tested in the future.

The methods of producing Subsurface images significatively improved the spatial prediction of soil properties. The technique can be applied for monitoring soil carbon in soil of mountainous regions of the Atlantic Forest Biome with very restricted access, such as the INP, where it can contribute to the park managers. It can also be used for agricultural purposes, and using of MID infrared might enhance the prediction results. The Subsurfacing process allows to validate each band of the spectral image, which permits the computation of uncertainties.

The Multispectral images provided great results for the HSM (close to Hyperspectral). So, whenever available, the use of data from the Hyperspectral sensor is recommended and newer sensors mighty provide even better results. In other words, for DSM or HSM, the Hyperspectral images can improve (slightly) the spatial predictions, but the strong gain comes with HSM in face of DSM. Although, it is recommended to test in other areas to confirm the strength of the technique, in this study the Subsurface Image outperformed the conventional covariates, consequently, the HSM outperformed the conventional DSM.

4 GENERAL CONCLUSIONS

The use of spectra of soil samples from the upper part of the Itatiaia National Park (INP) obtained through Proximal Remote Sensing (PSS) in comparison with laboratory chemical analyzes, allowed to reaffirm the potential of Remote Sensing techniques (RS) to predict soil properties.

In parallel to the PSS recording of soil from the upper part of the INP and the prediction of soil properties with Machine Learning (ML) algorithms, a new spectral preprocessing was developed, the Inverse of Reflectance to Factor of 10^4 (IRF4). When it was applied it increased the predictive capacity of certain ML models for soil properties such as Total Carbon (TC) and Hydrogen, when compared to commonly used preprocessing. The combination of IRF4 with established techniques has also raised the potential for predicting soil properties in many cases. This treatment is simple to apply and does not require adjustments for use.

The fusion between PSS and Multi–Hyperspectral images for Digital Mapping of the TC content in the soils of the INP produced a result considered excellent, with an increase in the predictive capacity by 75%. This combination, which is a novelty of this work, minimized the limitations of both the PSS and Digital Soil Mapping (DSM) techniques, and it was nominated Hyperspectral Soil Mapping (HSM). This junction allowed to amplify the radiometric resolution of hyperspectral images to the desired PSS radiometric resolution. In addition to mapping the TC content in the soil, it is assumed that this technique can be applied to other soil properties, so it is recommended to be tested in other environments.

The work with INP Soil–RS dataset together with open software like R, allowed to advance in the state of the art of RS techniques, for soil property prediction and also Digital Mapping of Soil Properties. As previously mentioned, the merge between PSS and DSM, coined by the author as HSM, is highly recommended for areas with difficult access–locomotion, due to the greater predictive capacity achieved. With fewer field collection points, it is possible to survey soil properties, which meets the previous demand required for Digital Soil Mapping. Therefore HSM reduces significantly survey costs of soil survey, with greater efficiency and agility with the same (or greater) reliability as the usual DSM, with application potential to environmental monitoring.

Following the development of RS techniques, with the spectral record accessible (with the PSS spectra in hands), both predictions with PSS and HSM can be further updated. The spectral data contains much information that is still little explored or not al all. Therefore, in the future, the INP spectral data may be reevaluated and reprocessed, thus serving as a reference (together with the park) for comparison and scientific improvement of the methods.

5 RECOMMENDATIONS AND FUTURE WORK

Aside from atmospheric and geometric corrections (with ESA provided software), all processing was conducted in FOSS (free open source software) environment. Programming languages and software such as R have a unique learning curve, but allow flexibility in dealing with information, files and varieties of ML algorithms. This allows fine adjustments, according to the user's determinations and objectives, which can lead to an increase in the quality and precision of the results, when compared to other computer programs, in general more complex and with less possibility of changes by the user. In this context, authors encourage the use and development in such platforms.

The following recommendations are made to future research in order to improve the capacity of the models to predict soil properties: i) Explore deeper the wavelengths (1000 - 2500 nm) to build the Subsurface image; ii) Continue to explore with MID infrared; iii) Application of covariates selection methods at both levels: prediction of the soil reflectance image (subsurface image) and, spatial prediction of the chosen soil property (HSM); iv) Test different (newer) hyperspectral sensors; v) Compute the cLHS with shadow treated images and compare the mapping results on the same points with a non shadow treated; vi) Use different machine learning algorithms for HSM, such as Artificial Neural Networks, Cubist, Keras, among others, with deeper tuning on each of these models; vii) Use a synthetic image as covariate for the subsurfacing process may increase capacity of HSM; viii) Compute the uncertainties on subsurfacing process and HSM; ix) Test the prediction in different areas and or environmental features such as phytophysiognomies and soil types or properties, we hypothesize that the DEM covariate would not be in top position when applying the HSM (which require subsurfacing process), as it happened in this study; x) Since the RapidEye was used as covariate for the subsurfacing with CHRIS, to run a completely independent test to compare Multi and Hyperspectral images in HSM; and xi) Apply these spectral techniques in different biomes and landscape conditions, or even in mountainous versus flat areas, in the same region, to assess whether the ability to predict the properties will vary with the relief.

6 REFERENCES

ADELINE, K. R. M.; GOMEZ, C.; GORRETTA, N.; ROGER, J.-M. Predictive ability of soil properties to spectral degradation from laboratory Vis-NIR spectroscopy data. 2017. Disponível em: <http://ac.els-cdn.com/S0016706116307637/1-s2.0-S0016706116307637-main.pdf?_tid=3b80cb16-471e-11e7-8588-00000aacb35d&acdnat=1496358263_fa77f7350b641a3f84221da6bb830181>.

ALVARES, C. A.; STAPE, J. L.; SENTELHAS, P. C.; GONÇALVES, J. L. d. M.; SPAROVEK, G. Köppen's climate classification map for Brazil. **Meteorologische Zeitschrift**, v. 22, n. 6, p. 711–728, 2013. ISSN 16101227. Disponível em: <https://doi.org/10.1127/0941-2948/2013/0507>.

ANTUNES, M. A. H.; DEBIASI, P.; COSTA, A. R. d.; GLERIANI, J. M. Correção atmosférica de imagens ALOS/AVINIR-2 utilizando o modelo 6S. **Revista Brasileira de Cartografia**, v. 4, n. 64, p. 531–539, 2012.

AXIMOFF, I. A.; ALVES, R. G.; RODRIGUES, R. d. C. Campos de Altitude do Itatiaia: Aspectos Ambientais, Biológico e Ecológicos. **Boletim do Parque Nacional do Itatiaia Nº 18**, p. 74, 2014.

BARRETO, C. G.; CAMPOS, J. B.; MENDES, R. D.; 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**, p. 215, 2013.

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 2. **Relatório Técnico Instituto Chico Mendes**, p. 117, 2013.

BASSER, J. R. Q. **LEARNING WITH CONTINUOUS CLASSES**. Sydney, Australia, 1992. 343–348 p. Disponível em: <https://sci2s.ugr.es/keel/pdf/algorithm/congreso/1992-Quinlan-AI.pdf>.

BECKWITH, R. S. Titration Curves of Soil Organic Matter. **Nature**, v. 184, n. 4687, p. 745–746, 8 1959. ISSN 0028-0836. Disponível em: <http://www.nature.com/articles/184745a0>.

BEGIEBING, S.; BACH, H. Analyses of hyperspectral and directional CHRIS data for agricultural monitoring using a canopy reflectance model. **European Space Agency, (Special Publication) ESA SP**, n. 578, p. 50–57, 2004. ISSN 03796566.

BEN-DOR, E.; CHABRILLAT, S.; DEMATTÊ, J. A. M.; TAYLOR, G. R.; HILL, J.; WHITING, M. L.; SOMMER, S. Using Imaging Spectroscopy to study soil properties. **Remote Sensing of Environment**, 2009. ISSN 00344257.

BEN-DOR, E.; PATKIN, K.; BANIN, A.; KARNIELI, A. Mapping of several soil properties using DAIS-7915 hyperspectral scanner data - A case study over soils in Israel. **International Journal of Remote Sensing**, v. 23, n. 6, p. 1043–1062, 2002. ISSN 01431161.

BOWERS, S. A.; HANKS, R. J. Reflection of Radiant Energy from Soils. v. 100, n. 2, 1964. Disponível em: <https://www.ars.usda.gov/ARSUserFiles/30200525/870Reflectionofradiantenergyfromsoils.pdf>.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001. ISSN 08856125.

BREIMAN, L.; CUTLER, A.; LIAW, A.; WIENER, M. **randomForest: Breiman and Cutler's Random Forests for Classification and Regression**. 2018. Disponível em: <https://cran.r-project.org/package=randomForest>.

Brockmann. **BEAM 5.0**. 2014. Disponível em: <http://www.brockmann-consult.de/cms/web/beam/welcome>.

CAMPBELL, J. B.; WYNNE, R. H. **Introduction to remote sensing**. 5. ed. London: The Guilford Press, 2011. 718 p. ISBN 978-1-60918-176-5.

CARR, D.; ported by Nicholas Lewin-Koh; MAECHLER, M.; contains copies of lattice functions written by Deepayan Sarkar. **hexbin: Hexagonal Binning Routines**. 2019. Disponível em: <https://cran.r-project.org/package=hexbin>.

CASTALDI, F.; PALOMBO, A.; SANTINI, F.; PASCUCCI, S.; PIGNATTI, S.; CASA, R. Evaluation of the potential of the current and forthcoming multispectral and hyperspectral imagers to estimate soil texture and organic carbon. **Remote Sensing of Environment**, v. 179, p. 54–65, 2016. ISSN 00344257. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0034425716301195>.

CEZAR, E.; NANNI, M. R.; GUERRERO, C.; JUNIOR, C. A. da S.; CRUCIOL, L. G. T.; CHICATI, M. L.; SILVA, G. F. C. Organic matter and sand estimates by spectroradiometry: Strategies for the development of models with applicability at a local scale. **Geoderma**, Elsevier, v. 340, p. 224–233, 4 2019. ISSN 0016-7061. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0016706117304482?via%3Dihub>.

CHAGAS, C. d. S.; PINHEIRO, H. S. K.; JUNIOR, W. d. C.; ANJOS, L. H. C. d.; PEREIRA, N. R.; BHERING, S. B. Data mining methods applied to map soil units on tropical hillslopes in Rio de Janeiro, Brazil. **Geoderma Regional**, Elsevier, v. 9, p. 47–55, 6 2017. ISSN 2352-0094. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2352009417300603?via%3Dihub>.

CHANG, C. W.; LAIRD, D. A. Near-infrared reflectance spectroscopic analysis of soil C and N. **Soil Science**, v. 167, n. 2, p. 110–116, 2002. ISSN 0038075X.

CHANG, C.-W.; LAIRD, D. A.; MAUSBACH, M. J.; HURBURGH, J. C. R. Near-Infrared Reflectance Spectroscopy–Principal Components Regression Analyses of Soil Properties. **Soil Science Society of America Journal**, p. 480–490, 2001.

CHICATI, M. S.; NANNI, M. R.; CHICATI, M. L.; FURLANETTO, R. H.; CEZAR, E.; OLIVEIRA, R. B. D. Hyperspectral remote detection as an alternative to correlate data of soil constituents. **Remote Sensing Applications: Society and Environment**, Elsevier, v. 16, p. 100270, 11 2019. ISSN 2352-9385. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2352938519302708?via%3Dihub>.

CLAIROTTE, M.; GRINAND, C.; KOUAKOUA, E.; THÉBAULT, A.; SABY, N. P.; BERNOUX, M.; BARTHÈS, B. G. National calibration of soil organic carbon concentration using diffuse infrared reflectance spectroscopy. **Geoderma**, 2016. ISSN 00167061.

CLARK, R. N. **Spectroscopy of rocks and minerals, and principles of spectroscopy**. 3. ed. Denver, Colorado: U.S. Geological Survey, 1999. v. 3. 3–58 p. ISSN 10869379. ISBN 0471294055. Disponível em: <https://pdfs.semanticscholar.org/1e02/a41e9ff88cdf80bbe41bd81ecde71e779548.pdf>.

COBLENTZ, W. W. Some Optical Properties of Iodine. III. **PHYSICAL LABORATORY OF CORNELL UNIVERSITY**, v. 1, p. 51–59, 1903. Disponível em: <https://journals.aps.org/>.

COBLENTZ, W. W. Preliminary Communication on the Infra-Red Absorption Spectra of Organic Compounds. **American Astronomical Society - Physical Laboratory, Cornell University**, v. 20, n. 1, p. 207, 1904. Disponível em: <http://articles.adsabs.harvard.edu/cgi-bin/nph-iarticle_query?1904ApJ....20..207C&defaultprint=YES&filetype=.pdf>.

COELHO, F. F.; GIASSON, E.; CAMPOS, A. R.; COSTA, J. J. F.; COBLINSKI, J. A.; FERREIRA, T. O. Digital soil class mapping in Brazil: a systematic review. **Sci. Agric. v**, v. 78, n. 5, p. 2021, 2019. Disponível em: <https://doi.org/10.1590/1678-992X-2019-0227>.

COLTHUP, N. B. Spectra-Structure Correlations in the Infra-Red Region. **Journal of the Optical Society Of America**, v. 40, n. 6, p. 397–400, 1950. Disponível em: <https://www.osapublishing.org/DirectPDFAccess/E850BF6E-9E7F-C383-00DCF127C8C2007E_50048/josa-40-6-397.pdf?da=1&id=50048&seq=0&mobile=no>.

CONFORTI, M.; CASTRIGNANÒ, A.; ROBUSTELLI, G.; SCARCIGLIA, F.; STELLUTI, M.; BUTTAFUOCO, G. Laboratory-based Vis–NIR spectroscopy and partial least square regression with spatially correlated errors for predicting spatial variation of soil organic matter content. **CATENA**, Elsevier, v. 124, p. 60–67, 1 2015. ISSN 0341-8162. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0341816214002537>.

COSTA, E. M.; PINHEIRO, H. S. K.; ANJOS, L. H. C. D.; MARCONDES, R. A. T.; GELSLEICHTER, Y. A. Mapping soil properties in a poorly-accessible area. **Revista Brasileira de Ciencia do Solo**, Rio de Janeiro, v. 44, p. 1–21, 2020. ISSN 18069657.

CUTTER, M. **HDFclean V2 software**. Guildford, UK: Surrey Research, 2006. Disponível em: <https://earth.esa.int/web/guest/-/hdfclean-v2-4409#_101_INSTANCE_cPK3_matmp>.

CUTTER, M.; KELLAR-BLAND, H. **CHRIS Data Format**. Guildford, UK: Surrey Satellite Technology Limited, 2008. (07/07/2008). Disponível em: <https://earth.esa.int/c/document_library/get_file?folderId=23844&name=DLFE-592.pdf>.

DANGAL, S.; SANDERMAN, J.; WILLS, S.; RAMIREZ-LOPEZ, L.; DANGAL, S. R. S.; SANDERMAN, J.; WILLS, S.; RAMIREZ-LOPEZ, L. Accurate and Precise Prediction of Soil Properties from a Large Mid-Infrared Spectral Library. **Soil Systems**, Multidisciplinary Digital Publishing Institute, v. 3, n. 1, p. 11, 1 2019. ISSN 2571-8789. Disponível em: <http://www.mdpi.com/2571-8789/3/1/11>.

DANIEL, K. W.; TRIPATHI, N. K.; HONDA, K. Artificial neural network analysis of laboratory and in situ spectra for the estimation of macronutrients in soils of Lop Buri (Thailand). **Soil Research**, v. 41, n. 1, p. 47, 2003. ISSN 1838-675X. Disponível em: <www.publish.csiro.au/journals/ajsrhttp://go.galegroup.com/ps/i.do?id=GALE%7CA98314097&sid=googleScholar&v=2.1&it=r&linkaccess=fulltext&issn=00049573&p=AONE&sw=w&authCount=1&u=nysl_ca_karigon&selfRedirect=truehttp://www.publish.csiro.au/?paper=SR02027>.

DAS, B. S.; RAY, S. S.; SARATHJITH, M. C.; SANTRA, P.; SAHOO, R. N.; SRIVASTAVA, R. Hyperspectral remote sensing: Opportunities, status and challenges for rapid soil assessment in India Hyperspectral remote sensing: opportunities, status and challenges for rapid soil assessment in India. n. September, 2015.

DEMATTÊ, J.; RAMIREZ-LOPEZ, L.; RIZZO, R.; NANNI, M.; FIORIO, P.; FONGARO, C.; NETO, L. M.; SAFANELLI, J.; BARROS, P. da S.; DEMATTÊ, J. A. M.; RAMIREZ-LOPEZ, L.; RIZZO, R.; NANNI, M. R.; FIORIO, P. R.; FONGARO, C. T.; NETO, L. G. M.; SAFANELLI, J. L.; BARROS, P. P. D. S. Remote Sensing from Ground to Space Platforms Associated with Terrain Attributes as a Hybrid Strategy on the Development of a Pedological Map. **Remote Sensing**, Multidisciplinary Digital Publishing Institute, v. 8, n. 10, p. 826, 10 2016. ISSN 2072-4292. Disponível em: <http://www.mdpi.com/2072-4292/8/10/826>.

DEMATTÊ, J. A.; DOTTO, A. C.; PAIVA, A. F.; SATO, M. V.; DALMOLIN, R. S.; ARAÚJO, M. d. S. B. de; SILVA, E. B. da; NANNI, M. R.; CATEN, A. ten; NORONHA, N. C.; LACERDA, M. P.; FILHO, J. C. de A.; RIZZO, R.; BELLINASO, H.; FRANCELINO, M. R.; SCHAEFER, C. E.; VICENTE, L. E.; SANTOS, U. J. dos; SAMPAIO, E. V. de S. B.; MENEZES, R. S.; SOUZA, J. J. L. de; ABRAHÃO, W. A.; COELHO, R. M.; GREGO, C. R.; LANI, J. L.; FERNANDES, A. R.; GONÇALVES, D. A.; SILVA, S. H.; MENEZES, M. D. de; CURI, N.; COUTO, E. G.; ANJOS, L. H. dos; CEDDIA, M. B.; PINHEIRO, E. F.; GRUNWALD, S.; VASQUES, G. M.; JÚNIOR, J. M.; SILVA, A. J. da; BARRETO, M. C. d. V.; NÓBREGA, G. N.; SILVA, M. Z. da; SOUZA, S. F. de; VALLADARES, G. S.; VIANA, J. H. M.; TERRA, F. da S.; HORÁK-TERRA, I.; FIORIO, P. R.; SILVA, R. C. da; JÚNIOR, E. F. F.; LIMA, R. H.; ALBA, J. M. F.; JUNIOR, V. S. de S.; BREFIN, M. D. L. M. S.; RUIVO, M. D. L. P.; FERREIRA, T. O.; BRAIT, M. A.; CAETANO, N. R.; BRINGHENTI, I.; MENDES, W. de S.; SAFANELLI, J. L.; GUIMARÃES, C. C.; POPPIEL, R. R.; SOUZA, A. B. e; QUESADA, C. A.; COUTO, H. T. Z. do. The Brazilian Soil Spectral Library (BSSL): A general view, application and challenges. **Geoderma**, Elsevier, p. 113793, 2019. ISSN 00167061. Disponível em: <https://doi.org/10.1016/j.geoderma.2019.05.043>.

DEMATTÊ, J. A. M. Characterization and discrimination of soils by their reflected electromagnetic energy. **Pesquisa Agropecuaria Brasileira**, v. 37, n. 10, p. 1445–1458, 2002. ISSN 0100204X.

DEMATTÊ, J. A. M.; FONGARO, C. T.; RIZZO, R.; SAFANELLI, J. L. Geospatial Soil Sensing System (GEOS3): A powerful data mining procedure to retrieve soil spectral reflectance from satellite images. **Remote Sensing of Environment**, Elsevier, v. 212, p. 161–175, 6 2018. ISSN 0034-4257. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0034425718302049?via%3Dihub>.

DEMATTÊ, J. A. M.; TERRA, F. d. S. Spectral pedology: A new perspective on evaluation of soils along pedogenetic alterations. **Geoderma**, 2013. Disponível em: <http://www.elsevier.com/authorsrights>.

DEMATTÊ, J. A. M.; TERRA, F. d. S.; QUARTAROLI, C. F. Spectral behavior of some modal soil profiles from São Paulo State, Brazil. **Bragantia**, v. 71, n. 3, p. 413–423, 2012. ISSN 1678-4499. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0006-87052012000300015&lng=en&nrm=iso&tlng=en>.

DONAGEMA, G. K.; CAMPOS, D. V. B. d.; CALDERANO, S. B.; TEIXEIRA, W. G.; VIANA, J. H. M. Manual de Métodos de Análise de Solos. n. October 2016, p. 230, 2011.

DOTTO, A. C.; DALMOLIN, R. S. D.; CATEN, A. ten; GRUNWALD, S. A systematic study on the application of scatter-corrective and spectral-derivative preprocessing for multivariate prediction of soil organic carbon by Vis-NIR spectra. **Geoderma**, 2018. ISSN 00167061.

FANG, Q.; HONG, H.; ZHAO, L.; KUKOLICH, S.; YIN, K.; WANG, C. Visible and Near-Infrared Reflectance Spectroscopy for Investigating Soil Mineralogy: A Review. **Journal of Spectroscopy**, v. 2018, p. 1–14, 2018. ISSN 2314-4920. Disponível em: <https://www.hindawi.com/journals/jspec/2018/3168974/>.

FORMAGGIO, A. R.; EPIPHANIO, J. C. N.; VALERIANO, M. M.; OLIVEIRA, J. B. Comportamento espectral (450-2.450 nm) de solos tropicais de São Paulo. v. 0, n. 2, p. 467–474, 1996.

FRITSCH, S.; GUENTHER, F.; WRIGHT, M. N. **neuralnet: Training of Neural Networks**. 2019. Disponível em: <https://cran.r-project.org/package=neuralnet>.

GALLO, B. C.; DEMATTÊ, J. A.; RIZZO, R.; SAFANELLI, J. L.; MENDES, W. d. S.; LEPSCH, I. F.; SATO, M. V.; ROMERO, D. J.; LACERDA, M. P. Multi-temporal satellite images on topsoil attribute quantification and the relationship with soil classes and geology. **Remote Sensing**, v. 10, n. 10, p. 21, 2018. ISSN 20724292.

GALVÃO, L. S.; PIZARRO, M. A.; EPIPHANIO, J. C. N. Variations in reflectance of tropical soils: Spectral-chemical composition relationships from AVIRIS data. **Remote Sensing of Environment**, v. 75, n. 2, p. 245–255, 2001. ISSN 00344257.

GALVÃO, L. S.; VITORELLO, I. Role of organic matter in obliterating the effects of iron on spectral reflectance and colour of Brazilian tropical soils. v. 19, n. 10, p. 1969–1979, 1998.

GHOLIZADEH, A.; BORůVKA, L.; SABERIOON, M.; VAŠÁT, R.; GHOLIZADEH, A.; BORůVKA, L.; SABERIOON, M.; VAŠÁT, R. A Memory-Based Learning Approach as Compared to Other Data Mining Algorithms for the Prediction of Soil Texture Using Diffuse Reflectance Spectra. **Remote Sensing**, Multidisciplinary Digital Publishing Institute, v. 8, n. 4, p. 341, 4 2016. ISSN 2072-4292. Disponível em: <http://www.mdpi.com/2072-4292/8/4/341>.

GOMES, L. C.; FARIA, R. M.; SOUZA, E. de; VELOSO, G. V.; SCHAEFER, C. E. G.; FILHO, E. I. F. Modelling and mapping soil organic carbon stocks in Brazil. **Geoderma**, Elsevier, v. 340, n. January, p. 337–350, 2019. ISSN 00167061. Disponível em: <https://doi.org/10.1016/j.geoderma.2019.01.007>.

GOMEZ, C.; LAGACHERIE, P.; COULOUMA, G. Regional predictions of eight common soil properties and their spatial structures from hyperspectral Vis–NIR data. **Geoderma**, v. 189-190, p. 176–185, 11 2012. ISSN 00167061. Disponível em: <http://s3.amazonaws.com/academia.edu.documents/43991751/Regional_predictions_of_eight_common_soi20160322-22365-3s8zy2.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1496365736&Signature=GWQVf7deHX9knvZrRJd3uSTe%2FhQ%3D&response-content-disposition=inline>.

GOMEZ, C.; ROSSEL, R. A. V.; MCBRATNEY, A. B. Soil organic carbon prediction by hyperspectral remote sensing and field vis-NIR spectroscopy: An Australian case study. **Geoderma**, v. 146, n. 3-4, p. 403–411, 2008. ISSN 00167061.

GUANTER, L.; ALONSO, L.; GÓMEZ-CHOVA, L.; MORENO, J. Algorithm Theoretical Basis Document: CHRIS/PROBA Atmospheric Correction Module. 2008. Disponível em: <http://www.brockmann-consult.de/beam-wiki/display/CBOX/CHRIS+Proba+Toolbox+for+BEAM>.

GUEVARA, M.; OLMEDO, G. F.; STELL, E.; YIGINI, Y.; DUARTE, Y. A.; HERNÁNDEZ, C. A.; ARÉVALO, G. E.; ARROYO-CRUZ, C. E.; BOLIVAR, A.; BUNNING, S.; CAÑAS, N. B.; CRUZ-GAISTARDO, C. O.; DAVILA, F.; ACQUA, M. D.; ENCINA, A.; TACONA, H. F.; FONTES, F.; HERRERA, J. A. H.; NAVARRO, A. R. I.; LOAYZA, V.; MANUELES, A. M.; JARA, F. M.; OLIVERA, C.; HERMOSILLA, R. O.; PEREIRA, G.; PRIETO, P.; RAMOS, I. A.; BRINA, J. C. R.; RIVERA, R.; RODRÍGUEZ-RODRÍGUEZ, J.; ROOPNARINE, R.; IBARRA, A. R.; RIVEIRO, K. A. R.; SCHULZ, G. A.; SPENCE, A.; VASQUES, G. M.; VARGAS, R. R.; VARGAS, R. No silver bullet for digital soil mapping: country-specific soil organic carbon estimates across Latin America. **SOIL**, v. 4, n. 3, p. 173–193, 8 2018. ISSN 2199-398X. Disponível em: <https://www.soil-journal.net/4/173/2018/>.

GUO, L.; ZHANG, H.; SHI, T.; CHEN, Y.; JIANG, Q.; LINDERMAN, M. Prediction of soil organic carbon stock by laboratory spectral data and airborne hyperspectral images. **Geoderma**, 2019. ISSN 00167061.

HIJMANS, R. J. **raster: Geographic Data Analysis and Modeling**. 2019. Disponível em: <https://cran.r-project.org/package=raster>.

HUNT, G. R. Spectral signatures of particulate minerals in the visible and near infrared. **Geophysics**, IO FIGS, v. 42, n. 3, p. 501–511, 1977. ISSN 0026-8976. Disponível em: <http://library.seg.org/https://courses.eas.ualberta.ca/eas451/hunt1977.pdf>.

HUNT, J. M.; TURNER, D. S. Determination of Mineral Constituents of Rocks by Infrared Spectroscopy. **Analytical Chemistry**, v. 25, n. 8, p. 1169–1174, 1953. ISSN 15206882.

HUNT, J. M.; WISHERD, M. P.; BONHAM, L. C. Infrared Absorption Spectra of Minerals and Other Inorganic Compounds. **Analytical Chemistry**, v. 22, n. 12, p. 1478–1497, 1950. ISSN 15206882.

IBGE. **Brazilian Institute of Geography and Statistics**. 2010. Disponível em: <https://www.ibge.gov.br/>.

IBGE. **Manual Técnico de Pedologia**. 3. ed. Rio de Janeiro: IBGE - Brazilian Institute of Geography and Statistics, 2015. v. 3. 430 p. ISSN 0103-9598. ISBN 9788524043598. Disponível em: <https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?id=295017&view=detalhes>.

IMBROISI, D.; GUARITÁ-SANTOS, A. J. M.; BARBOSA, S. S.; SHINTAKU, S. d. F.; MONTEIRO, H. J.; PONCE, G. A. E.; FURTADO, J. G.; TINOCO, C. J.; MELLO, D. C.; MACHADO, P. F. L. Gestão de resíduos químicos em universidades: Universidade de Brasília em foco. **Química Nova**, v. 29, n. 2, p. 404–409, 4 2006. ISSN 0100-4042. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-40422006000200037&lng=pt&nrm=iso&tlng=pt>.

IUSS Working Group WRB. **World Reference Base for Soil Resources**. 3. ed. Rome: FAO - FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED

NATIONS, 2015. v. 43. 264 p. ISSN 0014-4797. ISBN 9789251083697. Disponível em: <https://doi.org/10.1017/S0014479706394902>.

JABER, S. M.; LANT, C. L.; AL-QINNA, M. I. Estimating spatial variations in soil organic carbon using satellite hyperspectral data and map algebra. **International Journal of Remote Sensing**, v. 32, n. 18, p. 5077–5103, 2011. ISSN 0143-1161.

JENNY, H. **Factors of Soil Formation**. LWW, 1941. v. 52. 415 p. Disponível em: <https://journals.lww.com/soilsci/Citation/1941/11000/Factors_of_Soil_Formation.9.aspx>.

JENSEN, J. R. **Remote sensing of the environment: an earth resource perspective**. 2. ed. South Carolina: Pearson, 2014. 614 p. ISBN 9878560507061.

KOPAČKOVÁ, V.; BEN-DOR, E.; CARMON, N.; NOTESCO, G. Modelling Diverse Soil Attributes with Visible to Longwave Infrared Spectroscopy Using PLSR Employed by an Automatic Modelling Engine. **Remote Sensing**, v. 9, n. 2, p. 134, 2 2017. ISSN 2072-4292. Disponível em: <http://www.mdpi.com/2072-4292/9/2/134>.

KUANG, B.; TEKIN, Y.; MOUAZEN, A. M. Comparison between artificial neural network and partial least squares for on-line visible and near infrared spectroscopy measurement of soil organic carbon, pH and clay content. **Soil and Tillage Research**, Elsevier, v. 146, p. 243–252, 3 2015. ISSN 0167-1987. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167198714002475>.

KUHN, M.; QUINLAN, R. **Cubist: Rule- And Instance-Based Regression Modeling**. 2018. Disponível em: <https://cran.r-project.org/package=Cubist>.

LAGACHERIE, P.; MCBRATNEY, A.; VOLTZ, M. **Digital Soil Mapping: An Introductory Perspective**. 1. ed. Amsterdam: Elsevier, 2007. 658 p. ISSN 01662481. ISBN 9780080468075. Disponível em: <https://books.google.com.br/books?hl=pt-PT&lr=&id=OjhtrR5QgqMC&oi=fnd&pg=PP1&dq=Digital+Soil+Mapping:+an+introductory+perspective&ots=rGXyNqF5vl&sig=a96HWkO2-bZjpJ1jxzmAJ6e16ZY#v=onepage&q=DigitalSoilMapping%3Aanintroductoryperspective&f=falsehttps:>.

LAL, R. Soil health and carbon management. **Food and Energy Security**, n. 1, p. 1–11, 2016. ISSN 20483694. Disponível em: <http://doi.wiley.com/10.1002/fes3.96>.

LAMICHHANE, S.; KUMAR, L.; WILSON, B. Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: A review. **Geoderma**, Elsevier, v. 352, p. 395–413, 10 2019. ISSN 0016-7061. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0016706119300540>.

LAWRENCE, R. L.; WOOD, S. D.; SHELEY, R. L. Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (randomForest). **Remote Sensing of Environment**, v. 100, n. 3, p. 356–362, 2006. ISSN 00344257.

LIMA, L. A. d. S.; NEUMANN, M. R. B.; REATTO, A.; ROIG, H. L. Mapeamento de solos do tradicional ao digital. **Documentos 316 Embrapa Cerrados**, v. 316, n. Março, p. 52, 2013. Disponível em: <ainfo.cnptia.embrapa.br/digital/bitstream/item/116635/1/doc-316.pdf>.

LIU, Y.; SUN, X.; OUYANG, A. Nondestructive measurement of soluble solid content of navel orange fruit by visible–NIR spectrometric technique with PLSR and PCA-BPNN. **LWT - Food**

Science and Technology, Academic Press, v. 43, n. 4, p. 602–607, 5 2010. ISSN 0023-6438. Disponível em: <https://www.sciencedirect.com/science/article/pii/S002364380900303X>.

LU, G.; FEI, B. Medical hyperspectral imaging: a review. **Journal of biomedical optics**, v. 19, n. 1, p. 10901, 2014. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3895860&tool=pmcentrez&rendertype=abstract>.

MALEKI, M.; HOLM, L. V.; RAMON, H.; MERCKX, R.; BAERDEMAEKER, J. D.; MOUAZEN, A. Phosphorus Sensing for Fresh Soils using Visible and Near Infrared Spectroscopy. **Biosystems Engineering**, Academic Press, v. 95, n. 3, p. 425–436, 11 2006. ISSN 1537-5110. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1537511006002704>.

MALONE, B. **ithir: Soil data and some useful associated functions.** 2018.

MCBRATNEY, A. B.; SANTOS, M. L. M.; MINASNY, B. On digital soil mapping. **Geoderma**, Sydney, v. 117, n. 1-2, p. 3–52, 2003. ISSN 00167061.

MCGILL, W. J.; ROSSEL, R. A. V.; ZHEJIANG, Z. S. Improved estimates of organic carbon using proximally sensed vis – NIR spectra corrected by piecewise direct standardization. **European Journal of Soil Science**, 2015. Disponível em: <https://www.researchgate.net/profile/Wenjun_Ji/publication/277313637_Improved_estimates_of_organic_carbon_using_proximally_sensed_vis-NIR_spectra_corrected_by_piecewise_direct_standardization/links/5567687808aeccd77737890f.pdf>.

MENDES, W. d. S.; NETO, L. G. M.; DEMATTÊ, J. A.; GALLO, B. C.; RIZZO, R.; SAFANELLI, J. L.; FONGARO, C. T. Is it possible to map subsurface soil attributes by satellite spectral transfer models? **Geoderma**, Elsevier, v. 343, p. 269–279, 6 2019. ISSN 0016-7061. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0016706117321006?via%3Dihub>.

MEVIK, B.-H.; WEHRENS, R.; LILAND, K. H. **pls: Partial Least Squares and Principal Component Regression**. 2019. Disponível em: <https://cran.r-project.org/package=pls>.

MINASNY, B.; MCBRATNEY, A. B. A conditioned Latin hypercube method for sampling in the presence of ancillary information. **Computers & Geosciences**, v. 32, p. 1378–1388, 2006. Disponível em: <http://ac.els-cdn.com/S009830040500292X/1-s2.0-S009830040500292X-main.pdf?_tid=813b89b4-471b-11e7-9bf1-00000aacb35e&acdnat=1496357091_d2061d09a2d7ff9f1fb6c16bc645f076>.

MODENESI, M. Depósitos de vertente e evolução quaternária do planalto do Itatiaia. **Revista do Instituto Geológico**, v. 13, n. 1, p. 31–46, 1992. Disponível em: <http://turmalina.igc.usp.br/scielo.php?script=sci_abstract&amp;pid=S0100-929X1992000100002&amp;lng=pt&amp;nrm=iso&amp;tlng=en>.

MOREIRA, L. C. J.; TEIXEIRA, A. d. S.; GALVÃO, L. S. Potential of multispectral and hyperspectral data to detect saline-exposed soils in Brazil. **GIScience & Remote Sensing**, Taylor & Francis, v. 52, n. 4, p. 416–436, 7 2015. ISSN 1548-1603. Disponível em: <http://www.tandfonline.com/doi/full/10.1080/15481603.2015.1040227>.

MORELLOS, A.; PANTAZI, X.-E.; MOSHOU, D.; ALEXANDRIDIS, T.; WHETTON, R.; TZIOTZIOS, G.; WIEBENSOHN, J.; BILL, R. Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy. **Biosystems Engineering**, Academic Press, v. 152, p. 104–116, 12 2016. ISSN 1537-5110. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1537511015304165>.

MOUAZEN, A.; KUANG, B.; BAERDEMAEKER, J. D.; RAMON, H. Comparison among principal component, partial least squares and back propagation neural network analyses for accuracy of measurement of selected soil properties with visible and near infrared spectroscopy. **Geoderma**, Elsevier, v. 158, n. 1-2, p. 23–31, 8 2010. ISSN 0016-7061. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0016706110000753>.

MULDER, V. L.; BRUIN, S. de; SCHAEPMAN, M. E.; MAYR, T. R. The use of remote sensing in soil and terrain mapping - A review. **Geoderma**, Elsevier B.V., v. 162, n. 1-2, p. 1–19, 2011. ISSN 00167061. Disponível em: <http://dx.doi.org/10.1016/j.geoderma.2010.12.018>.

MURRELL, P. **gridBase: Integration of base and grid graphics**. 2014. Disponível em: <https://cran.r-project.org/package=gridBase>.

NANNI, M. R.; CEZAR, E.; JUNIOR, C. A. d. S.; SILVA, G. F. C.; GUALBERTO, A. A. da S. Partial least squares regression (PLSR) associated with spectral response to predict soil attributes in transitional lithologies. **Archives of Agronomy and Soil Science**, Taylor & Francis, v. 64, n. 5, p. 682–695, 4 2018. ISSN 0365-0340. Disponível em: <https://www.tandfonline.com/doi/full/10.1080/03650340.2017.1373185>.

NAWAR, S.; MOUAZEN, A.; NAWAR, S.; MOUAZEN, A. M. Comparison between Random Forests, Artificial Neural Networks and Gradient Boosted Machines Methods of On-Line Vis-NIR Spectroscopy Measurements of Soil Total Nitrogen and Total Carbon. **Sensors**, Multidisciplinary Digital Publishing Institute, v. 17, n. 10, p. 2428, 10 2017. ISSN 1424-8220. Disponível em: <http://www.mdpi.com/1424-8220/17/10/2428>.

NETO, O. R.; TEIXEIRA, A.; LEÃO, R.; MOREIRA, L.; GALVÃO, L. Hyperspectral Remote Sensing for Detecting Soil Salinization Using ProSpecTIR-VS Aerial Imagery and Sensor Simulation. **Remote Sensing**, Multidisciplinary Digital Publishing Institute, v. 9, n. 1, p. 42, 1 2017. ISSN 2072-4292. Disponível em: <http://www.mdpi.com/2072-4292/9/1/42>.

NEUWIRTH, E. **RColorBrewer: ColorBrewer Palettes**. 2014. Disponível em: <https://cran.r-project.org/package=RColorBrewer>.

NGUYEN, H.; BUI, X.-N.; TRAN, Q.-H.; MAI, N.-L. A new soft computing model for estimating and controlling blast-produced ground vibration based on Hierarchical K-means clustering and Cubist algorithms. **Applied Soft Computing**, Elsevier, v. 77, p. 376–386, 4 2019. ISSN 1568-4946. Disponível em: <https://www.sciencedirect.com/science/article/pii/S156849461930050X#b34>.

NOURI, M.; GOMEZ, C.; GORRETTA, N.; ROGER, J. M. Clay content mapping from airborne hyperspectral Vis-NIR data by transferring a laboratory regression model. 2017. Disponível em: <http://ac-els-cdn-com.ez30.periodicos.capes.gov.br/S0016706117303956/1-s2.0-S0016706117303956-main.pdf?_tid=752464bc-48b4-11e7-98b8-00000aab0f27&acdnat=1496532735_93b022d0e9e2022574527f4a9f254956>.

NUSSBAUM, M.; SPIESS, K.; BALTENSWEILER, A.; GROB, U.; KELLER, A.; GREINER, L.; SCHAEPMAN, M. E.; PAPRITZ, A. Evaluation of digital soil mapping approaches with large sets of environmental covariates. **SOIL**, Copernicus Publications on behalf of the European Geosciences Union, v. 4, n. 1, p. 1–22, 1 2018. ISSN 2199398X. Disponível em: <https://soil.copernicus.org/articles/4/1/2018/soil-4-1-2018.pdf>.

PADARIAN, J.; MINASNY, B.; MCBRATNEY, A. Using deep learning to predict soil properties from regional spectral data. **Geoderma Regional**, Elsevier, v. 16, p. e00198, 3 2019. ISSN 2352-0094. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2352009418302785>.

PADILHA, M. C. d. C.; VICENTE, L. E.; DEMATTÊ, J. A.; LOEBMANN, D. G. dos S. W.; VICENTE, A. K.; SALAZAR, D. F.; GUIMARÃES, C. C. B. Using Landsat and soil clay content to map soil organic carbon of oxisols and Ultisols near São Paulo, Brazil. **Geoderma Regional**, Elsevier, v. 21, p. e00253, 6 2020. ISSN 2352-0094. Disponível em: <https://www.sciencedirect.com/science/article/pii/S235200942030002X?via%3Dihub>.

PAIM, C.; PALMA, E.; EIFLER-LIMA, V. Gerenciar Resíduos Químicos:Uma Necessidade. **Caderno de Farmácia**, v. 18, p. 23–31, 2002. Disponível em: <https://analiticaqmcresiduos.paginas.ufsc.br/files/2013/10/farmacos-UFRGS.pdf>.

PEJOVIĆ, M.; NIKOLIĆ, M.; HEUVELINK, G. B.; HENGL, T.; KILIBARDA, M.; BAJAT, B. Sparse regression interaction models for spatial prediction of soil properties in 3D. **Computers and Geosciences**, v. 118, n. May, p. 1–13, 2018. ISSN 00983004.

PINHEIRO, E. F. M.; CEDDIA, M. B.; CLINGENSMITH, C. M.; GRUNWALD, S.; VASQUES, G. M. Prediction of Soil Physical and Chemical Properties by Visible and Near-Infrared Diffuse Reflectance Spectroscopy in the Central Amazon. **Remote Sensing**, Multidisciplinary Digital Publishing Institute, v. 9, n. 4, p. 293, 3 2017. ISSN 2072-4292. Disponível em: <http://www.mdpi.com/2072-4292/9/4/293>.

PINHEIRO, H. S. K.; BARBOSA, T. P. R.; ANTUNES, M. A. H.; CARVALHO, D. C. d.; NUMMER, A. R.; JUNIOR, W. d. C.; CHAGAS, C. d. S.; FERNANDES-FILHO, E. I.; PEREIRA, M. G. Assessment of Phytoecological Variability by Red-Edge Spectral Indices and Soil-Landscape Relationships. **Remote Sensing**, Multidisciplinary Digital Publishing Institute, v. 11, n. 20, p. 2448, 10 2019. ISSN 2072-4292. Disponível em: <https://www.mdpi.com/2072-4292/11/20/2448>.

QGIS Development Team. **QGIS Geographic Information System**. Open Source Geospatial Foundation Project, 2019. Disponível em: <http://qgis.osgeo.org>.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria: R Foundation for Statistical Computing, 2019. Disponível em: <https://www.r-project.org/>.

RapidEye. Satellite Imagery product specifications. 2012.

RIPLEY, B. **MASS: Support Functions and Datasets for Venables and Ripley's MASS**. 2019. Disponível em: <https://cran.r-project.org/package=MASS>.

ROSA, P. A. d. S.; RUBERTI, E. Nepheline syenites to syenites and granitic rocks of the Itatiaia Alkaline Massif, Southeastern Brazil: new geological insights into a migratory ring Complex. **Brazilian Journal of Geology**, Sociedade Brasileira de Geologia, v. 48, n. 2, p.

347–372, 5 2018. ISSN 2317-4692. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S2317-48892018000200347&lng=en&tlng=en>.

ROSENBLATT, F. The Perceptron: A probabilistic model for information storage and organization in the brain. **Psychological Review**, v. 65, n. 6, p. 386–408, 1958. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.335.3398&rep=rep1&type=pdf>.

ROSSEL, R. A. V. ParLeS: Software for chemometric analysis of spectroscopic data. 2008. Disponível em: <https://www.researchgate.net/profile/Raphael_Viscarra_Rossel/publication/223256650_ParLeS_Software_for_Chemometric_Analysis_of_Spectroscopic_Data/links/00b49537e8ad4322f1000000/ParLeS-Software-for-Chemometric-Analysis-of-Spectroscopic-Data.pdf>.

ROSSEL, R. A. V.; ADAMCHUK, V. I.; SUDDUTH, K. A.; MCKENZIE, N. J.; LOBSEY, C. **Proximal Soil Sensing. An Effective Approach for Soil Measurements in Space and Time**. 1. ed. Elsevier Inc., 2011. v. 113. 237–282 p. ISSN 00652113. ISBN 9780123864734. Disponível em: <http://dx.doi.org/10.1016/B978-0-12-386473-4.00010-5>.

ROSSEL, R. A. V.; CATTLE, S. R.; ORTEGA, A.; FOUAD, Y. In situ measurements of soil colour, mineral composition and clay content by vis–NIR spectroscopy. **Geoderma**, v. 150, p. 253–266, 2009. Disponível em: <https://djfextranet.agrsci.dk/sites/soilsensors/public/Documents/BoStenbergliterature/Rossel2009.pdf>.

ROSSEL, R. A. V.; JEON, Y. S.; ODEH, I. O. A.; MCBRATNEY, A. B. Using a legacy soil sample to develop a mid-IR spectral library. **Australian Journal of Soil Research**, 2008. ISSN 00049573.

ROSSEL, R. V.; BEHRENS, T. Using data mining to model and interpret soil diffuse reflectance spectra. **Geoderma**, Elsevier, v. 158, n. 1-2, p. 46–54, 8 2010. ISSN 00167061. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0016706109004315https://linkinghub.elsevier.com/retrieve/pii/S0016706109004315>.

ROSSEL, R. V.; WALVOORT, D.; MCBRATNEY, A.; JANIK, L.; SKJEMSTAD, J. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. **Geoderma**, v. 131, p. 59–75, 2005. Disponível em: <https://www.researchgate.net/profile/Raphael_Viscarra_Rossel/publication/223767466_Visible_Near_Infrared_Mid_Infrared_or_Combined_Diffuse_Reflectance_Spectroscopy_for_Simultaneous_Assessment_of_Various_Soil_Properties/links/54af47f10cf2b48e8ed63bf3.pdf>.

ROUDIER, P.; HEDLEY, C. B.; LOBSEY, C. R.; ROSSEL, R. A. V.; LEROUX, C. Evaluation of two methods to eliminate the effect of water from soil vis – NIR spectra for predictions of organic carbon. **Geoderma**, v. 296, p. 98–107, 2017. Disponível em: <https://www.researchgate.net/profile/Pierre_Roudier2/publication/314230371_Evaluation_of_two_methods_to_eliminate_the_effect_of_water_from_soil_vis-NIR_spectra_for_predictions_of_organic_carbon/links/58bdd87da6fdcc2d14eb4fe5/Evaluation-of-two-methods-to-e>.

Rulequest Research. **Rulequest, data mining with cubist**. 2019. Disponível em: <https://rulequest.com/cubist-info.html>.

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**, Elsevier B.V., v. 243-244, p. 214–227, 2015. ISSN 00167061. Disponível em: <http://dx.doi.org/10.1016/j.geoderma.2014.12.017>.

SANTOS, H. G. d.; JACOMINE, P. K. T.; ANJOS, L. H. C. d.; OLIVEIRA, V. A. d.; LUMBRERAS, J. F.; COELHO, M. R.; ALMEIDA, J. A. d.; FILHO, J. C. d. A.; OLIVEIRA, J. B. d.; CUNHA, T. J. F.; Embrapa Soils. **Brazilian soil classification system**. 5. ed. Brasília: Embrapa Soils, 2018. 303–343 p. ISBN 9788570358219.

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

SAVITZKY, A.; GOLAY, M. J. E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. **Analytical Chemistry**, American Chemical Society, v. 36, n. 8, p. 1627–1639, 7 1964. ISSN 0003-2700. Disponível em: <http://pubs.acs.org/doi/abs/10.1021/ac60214a047>.

SELIGE, T.; BÖHNER, J.; SCHMIDHALTER, U. High resolution topsoil mapping using hyperspectral image and field data in multivariate regression modeling procedures. **Geoderma**, Elsevier, v. 136, n. 1-2, p. 235–244, 12 2006. ISSN 0016-7061. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0016706106001030>.

SOARES, P. F. C.; ANJOS, L. H. C. dos; 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. 14, 2016. Disponível em: <https://www.rbcsjournal.org/pt-br/article/histosols-in-an-upper-montane-environment-in-the-itatiaia-plateau/>.

STEVENS, A.; RAMIREZ-LOPEZ, L. **prospectr: Miscellaneous functions for processing and sample selection of vis-NIR diffuse reflectance data**. 2014. Disponível em: <https://cran.r-project.org/package=prospectr>.

STEVENS, A.; WESEMAEL, B. V.; VANDENSCHRICK, G.; TOURÉ, S.; TYCHON, B. Detection of carbon stock change in agricultural soils using spectroscopic techniques. **Soil Science Society of America Journal**, v. 70, n. 3, p. 844–850, 2006. ISSN 03615995.

TANG, Y.; JONES, E.; MINASNY, B. Evaluating low-cost portable near infrared sensors for rapid analysis of soils from South Eastern Australia. **Geoderma Regional**, Elsevier, v. 20, p. e00240, 3 2020. ISSN 2352-0094. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2352009419302391?via%3Dihub>.

TEIXEIRA, P. C.; DONAGEMMA, G. K.; FONTANA, A.; TEIXEIRA, W. G. **Manual de metodos de analises**. 3. ed. Brasília: Embrapa, 2017. 574 p. ISBN 9788570357717.

TERRA, F. S.; DEMATTÊ, J. A.; ROSSEL, R. A. V. Spectral libraries for quantitative analyses of tropical Brazilian soils: Comparing vis–NIR and mid-IR reflectance data. **Geoderma**, Elsevier, v. 255-256, p. 81–93, 10 2015. ISSN 0016-7061. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0016706115001317>.

TERRA, F. S.; DEMATTÊ, J. A.; ROSSEL, R. A. V. Proximal spectral sensing in pedological assessments: vis–NIR spectra for soil classification based on weathering and pedogenesis. **Geoderma**, Elsevier, v. 318, p. 123–136, 5 2018. ISSN 0016-7061. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0016706117302768>.

TOMZHINSKI, G. W.; RIBEIRO, K. T.; FERNANDES, M. d. C. Análise geoecológica dos incêndios florestais do Parque Nacional do Itatiaia. **Boletim do Parque Nacional do Itatiaia N° 15**, p. 158, 2012.

TORGO, L. **DMwR: Functions and data for Data Mining with R**. 2013. Disponível em: <https://cran.r-project.org/package=DMwR>.

TZIACHRIS, P.; METAXA, E.; PAPADOPOULOS, F.; PAPADOPOULOU, M. Spatial Modelling and Prediction Assessment of Soil Iron Using Kriging Interpolation with pH as Auxiliary Information. **ISPRS International Journal of Geo-Information**, Multidisciplinary Digital Publishing Institute, v. 6, n. 9, p. 283, 9 2017. ISSN 2220-9964. Disponível em: <http://www.mdpi.com/2220-9964/6/9/283>.

USHEY, K.; ALLAIRE, J. J.; WICKHAM, H.; RITCHIE, G. **rstudioapi: Safely Access the RStudio API**. 2019. Disponível em: <https://cran.r-project.org/package=rstudioapi>.

VASQUES, G.; GRUNWALD, S.; SICKMAN, J. Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra. **Geoderma**, Elsevier, v. 146, n. 1-2, p. 14–25, 7 2008. ISSN 0016-7061. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0016706108000980#tblfn10>.

VASQUES, G. M.; DEMATTÊ, J. A. M.; ROSSEL, R. A. V.; RAMÍREZ-LÓPEZ, L.; TERRA, F. S. Soil classification using visible/near-infrared diffuse reflectance spectra from multiple depths. **Geoderma**, v. 223-225, p. 73–78, 2014. Disponível em: <http://ac.els-cdn.com/S0016706114000445/1-s2.0-S0016706114000445-main.pdf?_tid=4a1960be-4928-11e7-878a-00000aab0f27&acdnat=1496582484_99a5cc32c7256f6ede65092e6972dbda>.

VERMOTE, E. F.; TANRÉ, D.; DEUZÉ, J. L.; HERMAN, M.; MORCRETTE, J. 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. ISSN 01962892.

WADOUX, A.; MINASNY, B.; MCBRATNEY, A. Machine learning for digital soil mapping: applications, challenges and suggested solutions. **Earth-Science Reviews**, v. 210, 2020. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0012825220304050>.

WALKLEY, A.; BLACK, I. A. An Examination of the Degtjareff Method for Determining Soil Organic Matter, and a Proposed Modification of the Chromic Acid Titration Method. **Soil Science**, v. 37, n. 1, p. 29–38, 1 1934. ISSN 0038-075X. Disponível em: <http://journals.lww.com/00010694-193401000-00003>.

WANG, X.; CHEN, Y.; GUO, L.; LIU, L. Construction of the Calibration Set through Multivariate Analysis in Visible and Near-Infrared Prediction Model for Estimating Soil Organic Matter. **Remote Sensing**, Multidisciplinary Digital Publishing Institute, v. 9, n. 3, p. 201, 2 2017. ISSN 2072-4292. Disponível em: <http://www.mdpi.com/2072-4292/9/3/201>.

WICKHAM, H. **stringr: Simple, Consistent Wrappers for Common String Operations**. 2019. Disponível em: <https://cran.r-project.org/package=stringr>.

WICKHAM, H.; CHANG, W.; HENRY, L.; PEDERSEN, T. L.; TAKAHASHI, K.; WILKE, C.; WOO, K.; YUTANI, H. **ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics**. 2019. Disponível em: <https://cran.r-project.org/package=ggplot2>.

WICKHAM, H.; FRANÇOIS, R.; HENRY, L.; MÜLLER, K. **dplyr: A Grammar of Data Manipulation**. 2019. Disponível em: <https://cran.r-project.org/package=dplyr>.

WING, M. K.; 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.; WING, M. K. C. from J.; WESTON, S.; WILLIAMS, A.; KEEFER, C.; ENGELHARDT, A.; COOPER, T.; MAYER, Z.; KENKEL, B.; the R Core Team; BENESTY, M.; LESCARBEAU, R.; ZIEM, A.; SCRUCCA, L.; TANG, Y.; CANDAN, C.; HUNT., T. **caret: Classification and Regression Training**. 2019. Disponível em: <https://cran.r-project.org/package=caret>.

XIE, X.-L.; LI, A.-B. Improving spatial estimation of soil organic matter in a subtropical hilly area using covariate derived from vis-NIR spectroscopy. **Biosystems Engineering**, Academic Press, v. 152, p. 126–137, 12 2016. ISSN 1537-5110. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1537511015303639?dgcid=raven_sd_recommender_email>.

ZHANG, G. l.; LIU, F.; SONG, X. d. Recent progress and future prospect of digital soil mapping: A review. **Journal of Integrative Agriculture**, Nanjing, v. 16, n. 12, p. 2871–2885, 12 2017. ISSN 20953119. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2095311917617623>.

7 APPENDIX

APPENDIX A — Complete results of Chapter I

Complementary Table for Tables 2.4 and 2.5 of Chapter I.

Table S1: The 600 cross-validated groups of predicted soil properties across the models and pre-treatments.

Pre-treatments	Model	Soil property	R^2	MSE	RMSE	bias	RPD
AB-log	ann	Al	-0.484	2.631	1.592	0.063	0.944
AB-log	cb	Al	0.431	1.135	1.049	-0.079	1.405
AB-log	plsr	Al	0.382	1.159	1.072	0.031	1.364
AB-log	rf	Al	0.23	1.546	1.227	0.06	1.196
AB-log	ann	TC	0.025	27.364	5.058	0.428	1.154
AB-log	cb	TC	0.829	4.717	2.121	-0.14	2.652
AB-log	plsr	TC	0.819	5.003	2.19	-0.03	2.554
AB-log	rf	TC	0.769	6.714	2.527	0.036	2.213
AB-log	ann	H	0.091	0.473	0.679	-0.027	1.105
AB-log	cb	H	0.625	0.198	0.44	0.001	1.697
AB-log	plsr	H	0.565	0.228	0.471	0.007	1.598
AB-log	rf	H	0.53	0.24	0.486	0.003	1.543
AB-log	ann	N	0.491	0.049	0.208	-0.003	1.761
AB-log	cb	N	0.756	0.028	0.159	-0.005	2.203
AB-log	plsr	N	0.798	0.021	0.143	-0.003	2.37
AB-log	rf	N	0.691	0.032	0.177	0.002	1.935
AB-log	ann	Ca	-0.242	0.111	0.272	0.006	0.949
AB-log	cb	Ca	-0.224	0.088	0.259	-0.043	0.953
AB-log	plsr	Ca	-0.878	0.096	0.29	-0.002	0.787
AB-log	rf	Ca	-0.383	0.076	0.253	0.013	0.919
AB-log	ann	K	0.057	0.021	0.133	0	1.085
AB-log	cb	K	-0.837	0.025	0.136	-0.002	1.25
AB-log	plsr	K	-0.107	0.021	0.135	0.003	1.059
AB-log	rf	K	-0.109	0.022	0.141	0.011	0.998
AB-log	ann	Mg	-0.849	0.166	0.396	0.015	0.794
AB-log	cb	Mg	-0.135	0.101	0.314	-0.05	0.968
AB-log	plsr	Mg	-0.112	0.095	0.305	-0.002	1.002
AB-log	rf	Mg	-0.036	0.093	0.301	0.016	1.013
AB-log	ann	Na	-0.372	0.003	0.038	0	0.933
AB-log	cb	Na	-11.385	0.006	0.061	0	0.773
AB-log	plsr	Na	-3.472	0.004	0.054	0	0.594
AB-log	rf	Na	-1.31	0.003	0.044	0.003	0.874
AB-log	ann	P	-3.714	191.526	10.922	0.551	0.899
AB-log	cb	P	-0.354	85.214	8.583	-1.387	0.939
AB-log	plsr	P	-0.428	83.634	8.805	-0.076	0.869
AB-log	rf	P	-0.057	69.414	7.782	0.043	1.004
AB-log	ann	pH	-0.198	0.183	0.422	-0.011	0.957
AB-log	cb	pH	0.123	0.134	0.362	-0.014	1.115
AB-log	plsr	pH	0.123	0.133	0.362	-0.009	1.108
AB-log	rf	pH	0.095	0.138	0.369	0.005	1.081
CR	ann	Al	-0.063	2.126	1.438	-0.015	1.021
CR	cb	Al	0.381	1.198	1.078	-0.016	1.377
CR	plsr	Al	0.299	1.355	1.146	0.015	1.296
CR	rf	Al	0.388	1.225	1.097	0.062	1.322
CR	ann	TC	0.468	14.341	3.639	-0.07	1.576
CR	cb	TC	0.724	7.494	2.642	0.086	2.188
CR	plsr	TC	0.732	7.021	2.586	-0.05	2.214
CR	rf	TC	0.81	5.581	2.295	0.026	2.442
CR	ann	H	-0.116	0.551	0.736	-0.03	1.02
CR	cb	H	0.522	0.25	0.497	-0.021	1.492
CR	plsr	H	0.455	0.281	0.525	-0.002	1.422
CR	rf	H	0.497	0.265	0.51	0.001	1.46
CR	ann	N	0.642	0.039	0.192	0.009	1.799
CR	cb	N	0.625	0.038	0.19	0.006	1.849
CR	plsr	N	0.686	0.032	0.174	-0.003	2

Table S1 – continued on the next page

Table S1 – continued from previous page

Pre-treatments	Model	Soil property	R^2	MSE	RMSE	bias	RPD
CR	rf	N	0.755	0.026	0.157	0.004	2.166
CR	ann	Ca	-0.02	0.067	0.233	-0.002	1.012
CR	cb	Ca	-2.34	0.121	0.329	-0.024	0.744
CR	plsr	Ca	-1.748	0.113	0.324	0.001	0.7
CR	rf	Ca	-0.558	0.078	0.262	0.024	0.876
CR	ann	K	-0.136	0.023	0.144	0.002	0.962
CR	cb	K	-0.144	0.02	0.135	-0.009	1.047
CR	plsr	K	-0.066	0.02	0.135	0.002	1.024
CR	rf	K	0.249	0.016	0.116	0.009	1.28
CR	ann	Mg	0.047	0.086	0.29	-0.005	1.044
CR	cb	Mg	-0.364	0.123	0.341	-0.025	0.908
CR	plsr	Mg	-0.235	0.107	0.323	0.004	0.953
CR	rf	Mg	0.107	0.081	0.28	0.019	1.084
CR	ann	Na	-0.643	0.003	0.04	0	0.885
CR	cb	Na	-0.258	0.003	0.037	-0.005	0.99
CR	plsr	Na	-3.951	0.004	0.055	0	0.587
CR	rf	Na	-1.086	0.003	0.04	0.005	0.912
CR	ann	P	-0.165	79.209	8.24	0.177	0.959
CR	cb	P	-0.278	85.186	8.574	-1.374	0.92
CR	plsr	P	-0.702	96.102	9.427	0.189	0.824
CR	rf	P	-0.073	73.08	7.863	0.607	1.016
CR	ann	pH	-0.116	0.168	0.408	-0.002	0.977
CR	cb	pH	0.078	0.14	0.372	-0.018	1.073
CR	plsr	pH	-0.006	0.151	0.383	0	1.061
CR	rf	pH	0.18	0.124	0.351	0	1.133
IRF4	ann	Al	-0.286	2.393	1.518	0.159	0.99
IRF4	cb	Al	0.419	1.078	1.027	-0.078	1.449
IRF4	plsr	Al	0.176	1.67	1.268	0.043	1.168
IRF4	rf	Al	0.231	1.545	1.227	0.063	1.195
IRF4	ann	TC	0.524	14.238	3.624	-0.028	1.57
IRF4	cb	TC	0.852	3.998	1.958	-0.044	2.867
IRF4	plsr	TC	-0.701	60.745	5.01	0.006	1.84
IRF4	rf	TC	0.771	6.601	2.51	0.023	2.22
IRF4	ann	H	0.097	0.458	0.67	-0.043	1.132
IRF4	cb	H	0.672	0.173	0.411	-0.034	1.817
IRF4	plsr	H	-0.1	0.571	0.641	0.041	1.416
IRF4	rf	H	0.532	0.239	0.485	0.006	1.547
IRF4	ann	N	0.546	0.048	0.207	-0.003	1.75
IRF4	cb	N	0.731	0.03	0.161	-0.005	2.232
IRF4	plsr	N	0.375	0.066	0.23	0.006	1.731
IRF4	rf	N	0.685	0.033	0.179	0.003	1.908
IRF4	ann	Ca	-1.231	0.114	0.307	0.018	0.804
IRF4	cb	Ca	-0.307	0.091	0.266	-0.032	0.915
IRF4	plsr	Ca	-2.552	0.285	0.391	0.012	0.759
IRF4	rf	Ca	-0.374	0.076	0.253	0.013	0.923
IRF4	ann	K	-0.848	0.029	0.159	0.001	0.927
IRF4	cb	K	-0.551	0.023	0.13	-0.012	1.316
IRF4	plsr	K	-1.114	0.076	0.189	0.006	0.986
IRF4	rf	K	-0.117	0.022	0.141	0.011	0.993
IRF4	ann	Mg	-0.712	0.149	0.378	0.022	0.823
IRF4	cb	Mg	-0.056	0.095	0.304	-0.054	1.001
IRF4	plsr	Mg	-15.95	2.22	0.741	0.033	0.874
IRF4	rf	Mg	-0.037	0.093	0.301	0.017	1.013
IRF4	ann	Na	-1.626	0.003	0.041	-0.001	0.931
IRF4	cb	Na	-22.349	0.007	0.067	0	0.747
IRF4	plsr	Na	-102.221	0.039	0.118	0.006	0.474
IRF4	rf	Na	-1.345	0.003	0.044	0.003	0.864
IRF4	ann	P	-2.279	189.791	11.801	1.356	0.817
IRF4	cb	P	-0.335	80.372	8.382	-1.304	0.961
IRF4	plsr	P	-8.816	648.131	16.14	0.651	0.722
IRF4	rf	P	-0.062	70.088	7.812	0.052	1
IRF4	ann	pH	-0.328	0.204	0.44	-0.007	0.933
IRF4	cb	pH	0.22	0.117	0.34	-0.02	1.18
IRF4	plsr	pH	-2.228	0.457	0.558	-0.022	0.889
IRF4	rf	pH	0.096	0.138	0.369	0.006	1.081
IRF4 + NR 434	ann	Al	-0.01	1.816	1.329	0.04	1.131
IRF4 + NR 434	cb	Al	0.517	0.935	0.961	0.003	1.514
IRF4 + NR 434	plsr	Al	0.405	1.096	1.039	-0.012	1.431
IRF4 + NR 434	rf	Al	0.218	1.56	1.233	0.049	1.191

Table S1 – continued on the next page

Table S1 – continued from previous page

Pre-treatments	Model	Soil property	R^2	MSE	RMSE	bias	RPD
IRF4 + NR 434	ann	TC	0.472	12.351	3.401	0.366	1.728
IRF4 + NR 434	cb	TC	0.813	5.562	2.291	-0.204	2.451
IRF4 + NR 434	plsr	TC	0.824	4.965	2.181	-0.055	2.559
IRF4 + NR 434	rf	TC	0.774	6.521	2.493	0.048	2.24
IRF4 + NR 434	ann	H	0.066	0.47	0.672	-0.025	1.136
IRF4 + NR 434	cb	H	0.64	0.188	0.427	-0.026	1.766
IRF4 + NR 434	plsr	H	0.647	0.185	0.426	-0.019	1.745
IRF4 + NR 434	rf	H	0.498	0.257	0.502	0.005	1.503
IRF4 + NR 434	ann	N	0.624	0.04	0.186	-0.001	2.007
IRF4 + NR 434	cb	N	0.682	0.035	0.175	-0.001	2.054
IRF4 + NR 434	plsr	N	0.819	0.018	0.13	-0.005	2.649
IRF4 + NR 434	rf	N	0.697	0.032	0.175	0.002	1.952
IRF4 + NR 434	ann	Ca	-6.843	0.243	0.45	0.052	0.624
IRF4 + NR 434	cb	Ca	-0.704	0.094	0.276	-0.038	0.875
IRF4 + NR 434	plsr	Ca	-1.379	0.109	0.314	0.003	0.725
IRF4 + NR 434	rf	Ca	-0.345	0.075	0.251	0.012	0.934
IRF4 + NR 434	ann	K	-0.257	0.025	0.151	0.005	0.92
IRF4 + NR 434	cb	K	-0.95	0.027	0.15	-0.011	1.045
IRF4 + NR 434	plsr	K	-0.928	0.036	0.174	0.005	0.861
IRF4 + NR 434	rf	K	-0.109	0.022	0.141	0.012	0.999
IRF4 + NR 434	ann	Mg	-29.83	2.272	0.954	0.2	0.687
IRF4 + NR 434	cb	Mg	-0.143	0.101	0.315	-0.065	0.959
IRF4 + NR 434	plsr	Mg	-0.674	0.148	0.375	-0.007	0.828
IRF4 + NR 434	rf	Mg	-0.041	0.093	0.302	0.015	1.01
IRF4 + NR 434	ann	Na	-0.301	0.003	0.038	0	0.951
IRF4 + NR 434	cb	Na	-20.05	0.006	0.057	-0.003	0.902
IRF4 + NR 434	plsr	Na	-8.223	0.005	0.065	-0.001	0.505
IRF4 + NR 434	rf	Na	-1.328	0.003	0.044	0.003	0.854
IRF4 + NR 434	ann	P	-0.928	114.486	9.562	0.385	0.901
IRF4 + NR 434	cb	P	-0.071	75.301	7.954	-1.704	1.001
IRF4 + NR 434	plsr	P	-0.921	109.123	10.099	-0.02	0.763
IRF4 + NR 434	rf	P	-0.04	69.833	7.765	0.03	1.008
IRF4 + NR 434	ann	pH	-0.485	0.223	0.466	-0.024	0.873
IRF4 + NR 434	cb	pH	0.21	0.118	0.342	-0.007	1.172
IRF4 + NR 434	plsr	pH	-0.11	0.169	0.403	0	1.011
IRF4 + NR 434	rf	pH	0.081	0.14	0.372	0.005	1.071
IRF4 + SVG1-2-11	ann	Al	0.1	1.856	1.339	0.049	1.091
IRF4 + SVG1-2-11	cb	Al	0.398	1.166	1.056	-0.084	1.409
IRF4 + SVG1-2-11	plsr	Al	-8.592	21.427	2.541	-0.052	1.064
IRF4 + SVG1-2-11	rf	Al	0.527	0.967	0.965	0.039	1.527
IRF4 + SVG1-2-11	ann	TC	0.579	11.144	3.203	0.396	1.837
IRF4 + SVG1-2-11	cb	TC	0.819	4.896	2.167	-0.125	2.575
IRF4 + SVG1-2-11	plsr	TC	-2.989	148.945	6.607	-0.126	1.78
IRF4 + SVG1-2-11	rf	TC	0.841	4.753	2.112	-0.038	2.65
IRF4 + SVG1-2-11	ann	H	0.273	0.385	0.599	-0.015	1.321
IRF4 + SVG1-2-11	cb	H	0.602	0.207	0.447	0.006	1.702
IRF4 + SVG1-2-11	plsr	H	0.015	0.51	0.632	0.004	1.402
IRF4 + SVG1-2-11	rf	H	0.617	0.195	0.437	-0.004	1.724
IRF4 + SVG1-2-11	ann	N	0.544	0.053	0.208	0.008	1.897
IRF4 + SVG1-2-11	cb	N	0.741	0.027	0.161	-0.011	2.121
IRF4 + SVG1-2-11	plsr	N	-0.543	0.204	0.318	-0.001	1.604
IRF4 + SVG1-2-11	rf	N	0.812	0.02	0.138	-0.002	2.446
IRF4 + SVG1-2-11	ann	Ca	-2.12	0.129	0.319	0.016	0.809
IRF4 + SVG1-2-11	cb	Ca	-2.862	0.157	0.353	0.018	0.736
IRF4 + SVG1-2-11	plsr	Ca	-2.005	0.153	0.354	0.009	0.689
IRF4 + SVG1-2-11	rf	Ca	-0.471	0.079	0.261	0.021	0.881
IRF4 + SVG1-2-11	ann	K	0.084	0.02	0.13	-0.002	1.117
IRF4 + SVG1-2-11	cb	K	0.138	0.02	0.128	-0.007	1.144
IRF4 + SVG1-2-11	plsr	K	-1.611	0.079	0.204	0.005	0.908
IRF4 + SVG1-2-11	rf	K	0.215	0.017	0.12	0.006	1.24
IRF4 + SVG1-2-11	ann	Mg	-0.048	0.095	0.305	-0.001	0.996
IRF4 + SVG1-2-11	cb	Mg	-0.23	0.11	0.325	0.005	0.948
IRF4 + SVG1-2-11	plsr	Mg	-2.735	0.434	0.491	0.007	0.806
IRF4 + SVG1-2-11	rf	Mg	0.145	0.078	0.273	0.015	1.126
IRF4 + SVG1-2-11	ann	Na	-0.315	0.003	0.037	-0.001	0.96
IRF4 + SVG1-2-11	cb	Na	-19.047	0.007	0.064	0.002	0.737
IRF4 + SVG1-2-11	plsr	Na	-434.102	0.156	0.181	0.009	0.504
IRF4 + SVG1-2-11	rf	Na	-0.525	0.003	0.039	0.002	0.936
IRF4 + SVG1-2-11	ann	P	-2.677	151.246	10.146	1.297	0.92

Table S1 – continued on the next page

Table S1 – continued from previous page

Pre-treatments	Model	Soil property	R^2	MSE	RMSE	bias	RPD
IRF4 + SVG1-2-11	cb	P	-0.024	72.729	7.792	-1.313	1.019
IRF4 + SVG1-2-11	plsr	P	-9.319	673.771	16.989	0.786	0.663
IRF4 + SVG1-2-11	rf	P	0.032	68.528	7.548	0.468	1.072
IRF4 + SVG1-2-11	ann	pH	-0.111	0.171	0.408	0.015	0.988
IRF4 + SVG1-2-11	cb	pH	0.156	0.126	0.353	-0.015	1.138
IRF4 + SVG1-2-11	plsr	pH	-2.977	0.566	0.627	-0.012	0.785
IRF4 + SVG1-2-11	rf	pH	0.346	0.098	0.312	0	1.28
IRF4 + SVG-1-2-11 + NR 434	ann	Al	-0.05	2.094	1.432	0.019	1.018
IRF4 + SVG-1-2-11 + NR 434	cb	Al	0.464	1.112	1.037	-0.058	1.404
IRF4 + SVG-1-2-11 + NR 434	plsr	Al	0.263	1.387	1.165	-0.019	1.28
IRF4 + SVG-1-2-11 + NR 434	rf	Al	0.536	0.944	0.954	0.037	1.541
IRF4 + SVG-1-2-11 + NR 434	ann	TC	0.581	13.967	3.335	0.063	1.91
IRF4 + SVG-1-2-11 + NR 434	cb	TC	0.798	5.648	2.3	-0.212	2.461
IRF4 + SVG-1-2-11 + NR 434	plsr	TC	0.778	6.235	2.463	-0.065	2.229
IRF4 + SVG-1-2-11 + NR 434	rf	TC	0.84	4.749	2.113	-0.027	2.649
IRF4 + SVG-1-2-11 + NR 434	ann	H	0.198	0.418	0.632	-0.04	1.229
IRF4 + SVG-1-2-11 + NR 434	cb	H	0.627	0.194	0.436	-0.006	1.722
IRF4 + SVG-1-2-11 + NR 434	plsr	H	0.498	0.267	0.505	-0.014	1.509
IRF4 + SVG-1-2-11 + NR 434	rf	H	0.605	0.201	0.443	-0.002	1.699
IRF4 + SVG-1-2-11 + NR 434	ann	N	0.59	0.048	0.207	0.008	1.7
IRF4 + SVG-1-2-11 + NR 434	cb	N	0.764	0.024	0.15	-0.016	2.305
IRF4 + SVG-1-2-11 + NR 434	plsr	N	0.781	0.023	0.149	-0.005	2.267
IRF4 + SVG-1-2-11 + NR 434	rf	N	0.815	0.019	0.137	-0.002	2.466
IRF4 + SVG-1-2-11 + NR 434	ann	Ca	-0.457	0.088	0.266	0.011	0.905
IRF4 + SVG-1-2-11 + NR 434	cb	Ca	-1.401	0.121	0.319	0	0.761
IRF4 + SVG-1-2-11 + NR 434	plsr	Ca	-2.694	0.148	0.372	0.01	0.611
IRF4 + SVG-1-2-11 + NR 434	rf	Ca	-0.354	0.077	0.255	0.019	0.903
IRF4 + SVG-1-2-11 + NR 434	ann	K	0.076	0.02	0.131	-0.008	1.08
IRF4 + SVG-1-2-11 + NR 434	cb	K	-0.01	0.02	0.13	-0.006	1.138
IRF4 + SVG-1-2-11 + NR 434	plsr	K	-0.895	0.032	0.173	-0.001	0.806
IRF4 + SVG-1-2-11 + NR 434	rf	K	0.207	0.017	0.12	0.006	1.243
IRF4 + SVG-1-2-11 + NR 434	ann	Mg	-0.079	0.099	0.309	0.012	0.986
IRF4 + SVG-1-2-11 + NR 434	cb	Mg	-0.025	0.093	0.299	-0.014	1.027
IRF4 + SVG-1-2-11 + NR 434	plsr	Mg	-0.557	0.132	0.359	-0.009	0.855
IRF4 + SVG-1-2-11 + NR 434	rf	Mg	0.138	0.078	0.274	0.017	1.122
IRF4 + SVG-1-2-11 + NR 434	ann	Na	-0.582	0.003	0.039	0.001	0.898
IRF4 + SVG-1-2-11 + NR 434	cb	Na	-23.439	0.008	0.074	0.004	0.688
IRF4 + SVG-1-2-11 + NR 434	plsr	Na	-7.947	0.005	0.065	-0.001	0.511
IRF4 + SVG-1-2-11 + NR 434	rf	Na	-0.645	0.003	0.039	0.002	0.932
IRF4 + SVG-1-2-11 + NR 434	ann	P	-0.421	81.295	8.443	-0.064	0.941
IRF4 + SVG-1-2-11 + NR 434	cb	P	-0.078	71.454	7.796	-1.137	1.014
IRF4 + SVG-1-2-11 + NR 434	plsr	P	-1.348	122.642	10.846	-0.262	0.71
IRF4 + SVG-1-2-11 + NR 434	rf	P	0.02	69.58	7.589	0.552	1.075
IRF4 + SVG-1-2-11 + NR 434	ann	pH	-0.059	0.157	0.393	-0.015	1.023
IRF4 + SVG-1-2-11 + NR 434	cb	pH	0.183	0.122	0.347	-0.013	1.154
IRF4 + SVG-1-2-11 + NR 434	plsr	pH	-0.241	0.185	0.427	0.002	0.944
IRF4 + SVG-1-2-11 + NR 434	rf	pH	0.347	0.098	0.311	0	1.284
no pre-treatment	ann	Al	-0.285	2.529	1.557	0.059	0.969
no pre-treatment	cb	Al	0.362	1.264	1.111	-0.072	1.319
no pre-treatment	plsr	Al	0.115	1.64	1.261	-0.042	1.187
no pre-treatment	rf	Al	0.242	1.525	1.218	0.054	1.205
no pre-treatment	ann	TC	0.695	9.024	2.872	-0.095	2.021
no pre-treatment	cb	TC	0.824	5.128	2.153	-0.059	2.667
no pre-treatment	plsr	TC	0.727	6.63	2.548	-0.037	2.189
no pre-treatment	rf	TC	0.771	6.604	2.507	0.032	2.229
no pre-treatment	ann	H	0.27	0.374	0.605	0.031	1.243
no pre-treatment	cb	H	0.631	0.191	0.433	-0.023	1.721
no pre-treatment	plsr	H	0.592	0.212	0.459	0.004	1.611
no pre-treatment	rf	H	0.523	0.244	0.489	0.003	1.533
no pre-treatment	ann	N	0.596	0.041	0.198	0.005	1.769
no pre-treatment	cb	N	0.743	0.028	0.162	-0.009	2.141
no pre-treatment	plsr	N	0.676	0.03	0.172	-0.002	1.975
no pre-treatment	rf	N	0.69	0.032	0.177	0.002	1.917
no pre-treatment	ann	Ca	-3.928	0.143	0.352	0.028	0.747
no pre-treatment	cb	Ca	-0.401	0.092	0.27	-0.035	0.89
no pre-treatment	plsr	Ca	-1.192	0.095	0.297	0.005	0.757
no pre-treatment	rf	Ca	-0.4	0.075	0.253	0.014	0.925
no pre-treatment	ann	K	0.018	0.022	0.137	0.001	1.043
no pre-treatment	cb	K	-0.865	0.025	0.144	-0.004	1.103

Table S1 – continued on the next page

Table S1 – continued from previous page

Pre-treatments	Model	Soil property	R^2	MSE	RMSE	bias	RPD
no pre-treatment	plsr	K	-0.167	0.02	0.137	0	1.009
no pre-treatment	rf	K	-0.087	0.022	0.14	0.011	1.011
no pre-treatment	ann	Mg	-0.562	0.145	0.365	0.042	0.874
no pre-treatment	cb	Mg	-0.106	0.1	0.312	-0.065	0.97
no pre-treatment	plsr	Mg	-0.116	0.097	0.308	-0.005	0.989
no pre-treatment	rf	Mg	-0.031	0.093	0.301	0.018	1.015
no pre-treatment	ann	Na	-3.777	0.005	0.053	0.003	0.79
no pre-treatment	cb	Na	-10.222	0.006	0.062	0	0.836
no pre-treatment	plsr	Na	-2.819	0.003	0.05	0.001	0.65
no pre-treatment	rf	Na	-1.196	0.003	0.043	0.003	0.878
no pre-treatment	ann	P	-0.966	122.897	10.064	0.517	0.854
no pre-treatment	cb	P	-0.107	75.897	8.019	-1.422	0.993
no pre-treatment	plsr	P	-0.319	77.514	8.345	-0.152	0.938
no pre-treatment	rf	P	-0.051	69.888	7.781	0.048	1.006
no pre-treatment	ann	pH	-0.306	0.195	0.433	0.021	0.942
no pre-treatment	cb	pH	-0.033	0.156	0.393	-0.008	1.014
no pre-treatment	plsr	pH	-0.045	0.157	0.392	-0.008	1.031
no pre-treatment	rf	pH	0.096	0.139	0.37	0.004	1.08
PCAL	ann	Al	-0.029	2.061	1.406	0.004	1.067
PCAL	cb	Al	0.197	1.562	1.233	-0.023	1.203
PCAL	plsr	Al	0.057	1.724	1.293	-0.035	1.166
PCAL	rf	Al	0.249	1.516	1.218	0.056	1.203
PCAL	ann	TC	0.652	8.995	2.952	-0.061	1.8
PCAL	cb	TC	0.793	5.518	2.297	-0.148	2.338
PCAL	plsr	TC	0.717	6.24	2.488	-0.068	2.134
PCAL	rf	TC	0.757	6.528	2.496	0.039	2.146
PCAL	ann	H	0.189	0.394	0.625	-0.022	1.15
PCAL	cb	H	0.538	0.223	0.464	-0.026	1.596
PCAL	plsr	H	0.568	0.213	0.459	-0.001	1.565
PCAL	rf	H	0.521	0.235	0.481	0.004	1.509
PCAL	ann	N	0.423	0.054	0.22	0.02	1.645
PCAL	cb	N	0.701	0.031	0.173	-0.004	1.89
PCAL	plsr	N	0.667	0.029	0.17	-0.004	1.933
PCAL	rf	N	0.664	0.034	0.181	0.004	1.817
PCAL	ann	Ca	-0.777	0.094	0.281	0.007	0.87
PCAL	cb	Ca	-2.497	0.138	0.339	-0.016	0.765
PCAL	plsr	Ca	-1.417	0.101	0.306	0	0.741
PCAL	rf	Ca	-0.409	0.08	0.258	0.011	0.917
PCAL	ann	K	-0.929	0.045	0.187	0.02	0.803
PCAL	cb	K	-1.255	0.029	0.155	-0.005	1.067
PCAL	plsr	K	-0.193	0.021	0.138	-0.001	0.998
PCAL	rf	K	-0.143	0.023	0.143	0.011	0.973
PCAL	ann	Mg	-0.4	0.132	0.354	0.01	0.882
PCAL	cb	Mg	-0.103	0.101	0.314	-0.042	0.981
PCAL	plsr	Mg	-0.144	0.101	0.314	-0.005	0.982
PCAL	rf	Mg	-0.062	0.098	0.309	0.019	1
PCAL	ann	Na	-0.377	0.003	0.038	-0.001	0.967
PCAL	cb	Na	-11.993	0.005	0.058	0	0.79
PCAL	plsr	Na	-3.468	0.004	0.051	0.001	0.644
PCAL	rf	Na	-1.198	0.003	0.043	0.003	0.889
PCAL	ann	P	-0.681	80.899	8.103	0.019	0.938
PCAL	cb	P	-0.436	74.15	7.918	-1.156	0.927
PCAL	plsr	P	-0.433	71.914	7.927	-0.021	0.908
PCAL	rf	P	0.036	60.975	7.013	0.002	1.042
PCAL	ann	pH	-0.511	0.236	0.474	0.021	0.883
PCAL	cb	pH	0.097	0.141	0.373	-0.015	1.087
PCAL	plsr	pH	-0.089	0.169	0.406	-0.008	1.008
PCAL	rf	pH	0.091	0.143	0.375	0.006	1.075
RHCC	ann	Al	-0.298	2.596	1.555	0.137	0.985
RHCC	cb	Al	0.261	1.402	1.161	-0.076	1.295
RHCC	plsr	Al	0.137	1.598	1.245	0.012	1.201
RHCC	rf	Al	0.238	1.524	1.218	0.043	1.205
RHCC	ann	TC	0.755	7.184	2.58	-0.052	2.198
RHCC	cb	TC	0.789	6.179	2.41	-0.166	2.338
RHCC	plsr	TC	0.732	7.045	2.608	-0.067	2.146
RHCC	rf	TC	0.778	6.422	2.476	0.009	2.248
RHCC	ann	H	-0.046	0.533	0.722	0.002	1.036
RHCC	cb	H	0.537	0.241	0.485	-0.001	1.548
RHCC	plsr	H	0.479	0.277	0.519	-0.008	1.444

Table S1 – continued on the next page

Table S1 – continued from previous page

Pre-treatments	Model	Soil property	R^2	MSE	RMSE	bias	RPD
RHCC	rf	H	0.49	0.261	0.506	0.004	1.485
RHCC	ann	N	0.61	0.039	0.193	0.009	1.809
RHCC	cb	N	0.769	0.025	0.153	-0.008	2.265
RHCC	plsr	N	0.701	0.029	0.169	-0.004	1.995
RHCC	rf	N	0.694	0.032	0.176	0.001	1.935
RHCC	ann	Ca	-0.847	0.104	0.292	0.011	0.839
RHCC	cb	Ca	-0.624	0.079	0.259	-0.044	0.913
RHCC	plsr	Ca	-0.962	0.093	0.291	0.002	0.777
RHCC	rf	Ca	-0.345	0.075	0.252	0.009	0.922
RHCC	ann	K	-1.572	0.038	0.178	0.008	0.862
RHCC	cb	K	-0.515	0.023	0.136	-0.008	1.158
RHCC	plsr	K	-0.406	0.024	0.15	-0.001	0.923
RHCC	rf	K	-0.055	0.022	0.138	0.01	1.024
RHCC	ann	Mg	-0.908	0.149	0.374	0.017	0.862
RHCC	cb	Mg	-0.101	0.099	0.31	-0.059	0.98
RHCC	plsr	Mg	-0.01	0.09	0.296	0.001	1.031
RHCC	rf	Mg	-0.042	0.094	0.302	0.013	1.01
RHCC	ann	Na	-1.259	0.003	0.042	0.001	0.858
RHCC	cb	Na	-7.096	0.005	0.052	-0.002	0.903
RHCC	plsr	Na	-2.357	0.003	0.048	0.001	0.681
RHCC	rf	Na	-1.244	0.003	0.043	0.003	0.874
RHCC	ann	P	-0.286	83.166	8.412	0.254	0.96
RHCC	cb	P	-0.038	72.124	7.746	-1.41	1.04
RHCC	plsr	P	-0.341	83.167	8.569	0.053	0.916
RHCC	rf	P	-0.036	69.63	7.755	0.042	1.014
RHCC	ann	pH	-0.232	0.19	0.424	-0.012	0.969
RHCC	cb	pH	-0.037	0.157	0.394	-0.014	1.012
RHCC	plsr	pH	0.006	0.151	0.378	-0.011	1.096
RHCC	rf	pH	0.071	0.141	0.374	0.003	1.067
stepAIC	ann	Al	-0.191	2.223	1.445	0.16	1.064
stepAIC	cb	Al	0.423	1.152	1.062	-0.117	1.371
stepAIC	plsr	Al	0.253	1.364	1.157	0.015	1.29
stepAIC	rf	Al	0.226	1.553	1.228	0.062	1.197
stepAIC	ann	TC	0.591	10.659	3.041	-0.264	2.037
stepAIC	cb	TC	0.803	5.339	2.266	-0.052	2.466
stepAIC	plsr	TC	0.695	7.5	2.713	-0.037	2.051
stepAIC	rf	TC	0.778	6.502	2.485	0.045	2.246
stepAIC	ann	H	0.315	0.345	0.58	0.034	1.309
stepAIC	cb	H	0.594	0.211	0.454	0	1.648
stepAIC	plsr	H	0.523	0.245	0.49	0.003	1.53
stepAIC	rf	H	0.525	0.243	0.489	0.004	1.532
stepAIC	ann	N	0.666	0.034	0.18	0.008	1.932
stepAIC	cb	N	0.728	0.028	0.164	-0.009	2.085
stepAIC	plsr	N	0.636	0.033	0.181	-0.001	1.884
stepAIC	rf	N	0.706	0.031	0.173	0.003	1.982
stepAIC	ann	Ca	-1.728	0.109	0.311	0.021	0.791
stepAIC	cb	Ca	-0.162	0.074	0.245	-0.046	0.965
stepAIC	plsr	Ca	-0.854	0.087	0.282	0.008	0.796
stepAIC	rf	Ca	-0.516	0.078	0.26	0.015	0.896
stepAIC	ann	K	-0.972	0.041	0.185	0.017	0.8
stepAIC	cb	K	-0.067	0.023	0.133	-0.005	1.194
stepAIC	plsr	K	0.017	0.019	0.13	-0.001	1.065
stepAIC	rf	K	-0.07	0.022	0.139	0.01	1.008
stepAIC	ann	Mg	-1.79	0.209	0.426	0.04	0.799
stepAIC	cb	Mg	-0.038	0.093	0.301	-0.056	1.007
stepAIC	plsr	Mg	-0.062	0.094	0.302	0.006	1.01
stepAIC	rf	Mg	-0.065	0.096	0.305	0.015	0.996
stepAIC	ann	Na	-7.165	0.007	0.059	0.005	0.796
stepAIC	cb	Na	-7.107	0.005	0.052	-0.001	0.916
stepAIC	plsr	Na	-2.888	0.003	0.05	0	0.657
stepAIC	rf	Na	-1.593	0.003	0.045	0.003	0.885
stepAIC	ann	P	-0.931	94.434	8.908	0.019	0.945
stepAIC	cb	P	0	70.808	7.719	-1.66	1.026
stepAIC	plsr	P	-0.416	80.903	8.533	-0.063	0.923
stepAIC	rf	P	-0.056	69.372	7.779	0.098	1.005
stepAIC	ann	pH	-0.43	0.218	0.462	-0.008	0.872
stepAIC	cb	pH	0.014	0.151	0.385	-0.001	1.042
stepAIC	plsr	pH	0.079	0.14	0.371	-0.015	1.086
stepAIC	rf	pH	0.097	0.137	0.369	0.004	1.082

Table S1 – continued on the next page

Table S1 – continued from previous page

Pre-treatments	Model	Soil property	R^2	MSE	RMSE	bias	RPD
SVG-1-2-11	ann	Al	0.021	1.983	1.379	0.059	1.075
SVG-1-2-11	cb	Al	-0.05	1.994	1.371	0.042	1.106
SVG-1-2-11	plsr	Al	0.137	1.598	1.243	-0.027	1.209
SVG-1-2-11	rf	Al	0.513	0.987	0.974	0.049	1.52
SVG-1-2-11	ann	TC	0.553	12.93	3.348	0.184	1.884
SVG-1-2-11	cb	TC	0.734	7.393	2.655	0.085	2.121
SVG-1-2-11	plsr	TC	0.637	8.938	2.944	-0.11	1.911
SVG-1-2-11	rf	TC	0.836	4.718	2.123	0	2.627
SVG-1-2-11	ann	H	0.11	0.458	0.668	0.022	1.14
SVG-1-2-11	cb	H	0.278	0.367	0.6	-0.009	1.252
SVG-1-2-11	plsr	H	0.495	0.264	0.511	-0.021	1.45
SVG-1-2-11	rf	H	0.587	0.214	0.459	0.008	1.619
SVG-1-2-11	ann	N	0.546	0.039	0.191	0.009	1.893
SVG-1-2-11	cb	N	0.643	0.037	0.189	-0.012	1.826
SVG-1-2-11	plsr	N	0.589	0.039	0.198	-0.008	1.706
SVG-1-2-11	rf	N	0.797	0.021	0.142	0.003	2.382
SVG-1-2-11	ann	Ca	-1.667	0.086	0.264	0	0.964
SVG-1-2-11	cb	Ca	-3.158	0.139	0.359	-0.003	0.685
SVG-1-2-11	plsr	Ca	-1.343	0.103	0.309	0.014	0.736
SVG-1-2-11	rf	Ca	-0.498	0.082	0.264	0.021	0.882
SVG-1-2-11	ann	K	-0.094	0.022	0.14	0.003	0.991
SVG-1-2-11	cb	K	0.174	0.02	0.127	-0.012	1.153
SVG-1-2-11	plsr	K	-0.142	0.019	0.135	0	1.013
SVG-1-2-11	rf	K	0.192	0.017	0.12	0.007	1.17
SVG-1-2-11	ann	Mg	-0.347	0.117	0.335	0.013	0.923
SVG-1-2-11	cb	Mg	-0.286	0.114	0.332	-0.021	0.925
SVG-1-2-11	plsr	Mg	-0.154	0.101	0.315	0.006	0.966
SVG-1-2-11	rf	Mg	0.194	0.074	0.267	0.014	1.148
SVG-1-2-11	ann	Na	-0.35	0.003	0.038	0	0.95
SVG-1-2-11	cb	Na	-1.446	0.003	0.04	-0.003	0.931
SVG-1-2-11	plsr	Na	-7.089	0.005	0.062	0.001	0.516
SVG-1-2-11	rf	Na	-1.357	0.003	0.04	0.004	0.905
SVG-1-2-11	ann	P	-0.104	78.887	8.037	-0.13	1.001
SVG-1-2-11	cb	P	-0.14	78.031	8.118	-1.761	0.985
SVG-1-2-11	plsr	P	-1.088	109.271	10.155	-0.323	0.762
SVG-1-2-11	rf	P	0.021	67.732	7.548	0.471	1.051
SVG-1-2-11	ann	pH	0.062	0.143	0.376	-0.002	1.063
SVG-1-2-11	cb	pH	0.042	0.144	0.378	-0.015	1.053
SVG-1-2-11	plsr	pH	-0.064	0.158	0.395	-0.004	1.016
SVG-1-2-11	rf	pH	0.363	0.096	0.309	-0.005	1.286
SVG-1-2-11 + IRF4	ann	Al	-0.245	2.44	1.531	0.017	0.972
SVG-1-2-11 + IRF4	cb	Al	0.234	1.571	1.229	-0.097	1.2
SVG-1-2-11 + IRF4	plsr	Al	-84.456	141.554	7.007	0.567	0.462
SVG-1-2-11 + IRF4	rf	Al	0.487	1.071	1.012	0.034	1.459
SVG-1-2-11 + IRF4	ann	TC	0.377	13.865	3.596	0.149	1.694
SVG-1-2-11 + IRF4	cb	TC	0.736	7.778	2.694	0.163	2.082
SVG-1-2-11 + IRF4	plsr	TC	-55.531	1742.886	26.977	1.488	0.381
SVG-1-2-11 + IRF4	rf	TC	0.826	5.047	2.194	-0.013	2.539
SVG-1-2-11 + IRF4	ann	H	0.078	0.48	0.689	-0.042	1.077
SVG-1-2-11 + IRF4	cb	H	-0.001	0.517	0.706	0.004	1.077
SVG-1-2-11 + IRF4	plsr	H	-35.454	20.454	3.083	0.356	0.439
SVG-1-2-11 + IRF4	rf	H	0.568	0.224	0.47	0.003	1.58
SVG-1-2-11 + IRF4	ann	N	0.4	0.058	0.232	0.008	1.549
SVG-1-2-11 + IRF4	cb	N	0.61	0.045	0.203	-0.016	1.721
SVG-1-2-11 + IRF4	plsr	N	-57.865	7.296	1.681	0.094	0.381
SVG-1-2-11 + IRF4	rf	N	0.787	0.022	0.146	0	2.321
SVG-1-2-11 + IRF4	ann	Ca	-0.715	0.084	0.268	-0.004	0.891
SVG-1-2-11 + IRF4	cb	Ca	-2.28	0.162	0.329	-0.009	0.849
SVG-1-2-11 + IRF4	plsr	Ca	-36.225	0.955	0.675	-0.041	0.515
SVG-1-2-11 + IRF4	rf	Ca	-0.168	0.073	0.244	0.007	0.958
SVG-1-2-11 + IRF4	ann	K	-0.237	0.023	0.143	-0.006	1.004
SVG-1-2-11 + IRF4	cb	K	-0.421	0.028	0.156	-0.008	0.918
SVG-1-2-11 + IRF4	plsr	K	-30.568	0.577	0.594	-0.038	0.34
SVG-1-2-11 + IRF4	rf	K	0.274	0.017	0.118	0.003	1.242
SVG-1-2-11 + IRF4	ann	Mg	-0.329	0.119	0.338	-0.005	0.907
SVG-1-2-11 + IRF4	cb	Mg	-0.439	0.127	0.35	-0.023	0.879
SVG-1-2-11 + IRF4	plsr	Mg	-13.912	1.212	0.897	-0.098	0.49
SVG-1-2-11 + IRF4	rf	Mg	0.135	0.078	0.276	0.002	1.098
SVG-1-2-11 + IRF4	ann	Na	-0.507	0.003	0.04	-0.001	0.888

Table S1 – continued on the next page

Table S1 – continued from previous page

Pre-treatments	Model	Soil property	R^2	MSE	RMSE	bias	RPD
SVG-1-2-11 + IRF4	cb	Na	-19.333	0.006	0.061	0.001	0.777
SVG-1-2-11 + IRF4	plsr	Na	-202.6	0.097	0.233	0.022	0.281
SVG-1-2-11 + IRF4	rf	Na	-0.4	0.003	0.038	0.001	0.965
SVG-1-2-11 + IRF4	ann	P	-0.341	86.279	8.572	0.011	0.939
SVG-1-2-11 + IRF4	cb	P	-1.191	133.208	10.364	-0.715	0.828
SVG-1-2-11 + IRF4	plsr	P	-14.752	820.441	24.617	-0.47	0.414
SVG-1-2-11 + IRF4	rf	P	0.06	67.103	7.467	0.158	1.063
SVG-1-2-11 + IRF4	ann	pH	-0.137	0.176	0.412	-0.004	0.979
SVG-1-2-11 + IRF4	cb	pH	-0.131	0.17	0.408	-0.008	0.992
SVG-1-2-11 + IRF4	plsr	pH	-70.768	11.581	1.924	0.026	0.458
SVG-1-2-11 + IRF4	rf	pH	0.322	0.102	0.319	-0.005	1.249
SVG-1-2-11 + IRF4 + NR 434	ann	Al	0.033	1.98	1.376	-0.015	1.08
SVG-1-2-11 + IRF4 + NR 434	cb	Al	0.002	1.92	1.355	-0.056	1.105
SVG-1-2-11 + IRF4 + NR 434	plsr	Al	-84.671	141.719	6.988	0.565	0.463
SVG-1-2-11 + IRF4 + NR 434	rf	Al	0.506	1.028	0.992	0.032	1.485
SVG-1-2-11 + IRF4 + NR 434	ann	TC	0.501	13.818	3.609	0.092	1.586
SVG-1-2-11 + IRF4 + NR 434	cb	TC	0.72	8.274	2.788	0.049	2.019
SVG-1-2-11 + IRF4 + NR 434	plsr	TC	-56.013	1760.657	27.079	1.458	0.379
SVG-1-2-11 + IRF4 + NR 434	rf	TC	0.824	5.006	2.183	0.014	2.569
SVG-1-2-11 + IRF4 + NR 434	ann	H	-0.016	0.513	0.707	0.011	1.068
SVG-1-2-11 + IRF4 + NR 434	cb	H	0.131	0.433	0.647	0.054	1.179
SVG-1-2-11 + IRF4 + NR 434	plsr	H	-36.175	20.882	3.112	0.356	0.434
SVG-1-2-11 + IRF4 + NR 434	rf	H	0.541	0.239	0.486	-0.001	1.521
SVG-1-2-11 + IRF4 + NR 434	ann	N	0.193	0.085	0.285	0.02	1.218
SVG-1-2-11 + IRF4 + NR 434	cb	N	0.641	0.04	0.188	-0.01	1.948
SVG-1-2-11 + IRF4 + NR 434	plsr	N	-58.464	7.385	1.687	0.093	0.38
SVG-1-2-11 + IRF4 + NR 434	rf	N	0.777	0.022	0.148	0.002	2.318
SVG-1-2-11 + IRF4 + NR 434	ann	Ca	-2.08	0.123	0.324	0.028	0.788
SVG-1-2-11 + IRF4 + NR 434	cb	Ca	-1.642	0.14	0.321	-0.005	0.82
SVG-1-2-11 + IRF4 + NR 434	plsr	Ca	-36.106	0.951	0.674	-0.041	0.516
SVG-1-2-11 + IRF4 + NR 434	rf	Ca	-0.142	0.072	0.243	0.006	0.965
SVG-1-2-11 + IRF4 + NR 434	ann	K	-0.317	0.025	0.151	-0.007	0.919
SVG-1-2-11 + IRF4 + NR 434	cb	K	-0.61	0.031	0.165	-0.003	0.863
SVG-1-2-11 + IRF4 + NR 434	plsr	K	-31.138	0.587	0.598	-0.039	0.338
SVG-1-2-11 + IRF4 + NR 434	rf	K	0.275	0.017	0.118	0.003	1.244
SVG-1-2-11 + IRF4 + NR 434	ann	Mg	-0.217	0.105	0.321	-0.006	0.948
SVG-1-2-11 + IRF4 + NR 434	cb	Mg	-0.425	0.125	0.346	-0.029	0.892
SVG-1-2-11 + IRF4 + NR 434	plsr	Mg	-14.048	1.224	0.9	-0.097	0.488
SVG-1-2-11 + IRF4 + NR 434	rf	Mg	0.128	0.079	0.277	0.003	1.093
SVG-1-2-11 + IRF4 + NR 434	ann	Na	-0.601	0.003	0.039	0.001	0.892
SVG-1-2-11 + IRF4 + NR 434	cb	Na	-13.51	0.004	0.05	0	0.85
SVG-1-2-11 + IRF4 + NR 434	plsr	Na	-200.571	0.096	0.232	0.021	0.287
SVG-1-2-11 + IRF4 + NR 434	rf	Na	-0.341	0.003	0.038	0.001	0.971
SVG-1-2-11 + IRF4 + NR 434	ann	P	-0.431	94.117	8.963	-0.123	0.896
SVG-1-2-11 + IRF4 + NR 434	cb	P	-0.948	118.427	9.781	-0.759	0.859
SVG-1-2-11 + IRF4 + NR 434	plsr	P	-14.921	831.006	24.737	-0.526	0.412
SVG-1-2-11 + IRF4 + NR 434	rf	P	0.072	66.896	7.436	0.137	1.07
SVG-1-2-11 + IRF4 + NR 434	ann	pH	-0.139	0.172	0.413	0.013	0.963
SVG-1-2-11 + IRF4 + NR 434	cb	pH	-0.121	0.172	0.408	-0.003	0.992
SVG-1-2-11 + IRF4 + NR 434	plsr	pH	-73.21	11.976	1.946	0.026	0.456
SVG-1-2-11 + IRF4 + NR 434	rf	pH	0.329	0.101	0.317	-0.005	1.255
SVG-1-2-11 + NR 434	ann	Al	0.264	1.518	1.203	-0.01	1.233
SVG-1-2-11 + NR 434	cb	Al	0.019	1.915	1.338	-0.022	1.13
SVG-1-2-11 + NR 434	plsr	Al	0.107	1.675	1.271	-0.028	1.182
SVG-1-2-11 + NR 434	rf	Al	0.522	0.97	0.966	0.042	1.529
SVG-1-2-11 + NR 434	ann	TC	0.508	12.849	3.335	0.433	1.804
SVG-1-2-11 + NR 434	cb	TC	0.712	8.06	2.775	-0.008	2.022
SVG-1-2-11 + NR 434	plsr	TC	0.66	8.878	2.933	-0.036	1.91
SVG-1-2-11 + NR 434	rf	TC	0.831	4.854	2.143	0.024	2.632
SVG-1-2-11 + NR 434	ann	H	0.123	0.466	0.676	0.045	1.102
SVG-1-2-11 + NR 434	cb	H	0.178	0.43	0.649	-0.016	1.15
SVG-1-2-11 + NR 434	plsr	H	0.486	0.27	0.513	-0.013	1.456
SVG-1-2-11 + NR 434	rf	H	0.574	0.222	0.468	0.005	1.581
SVG-1-2-11 + NR 434	ann	N	0.513	0.049	0.212	-0.003	1.696
SVG-1-2-11 + NR 434	cb	N	0.621	0.04	0.196	-0.015	1.749
SVG-1-2-11 + NR 434	plsr	N	0.607	0.039	0.195	-0.006	1.752
SVG-1-2-11 + NR 434	rf	N	0.786	0.021	0.144	0.005	2.375
SVG-1-2-11 + NR 434	ann	Ca	-0.144	0.091	0.256	0.006	0.966
SVG-1-2-11 + NR 434	cb	Ca	-3.173	0.139	0.359	-0.003	0.685

Table S1 – continued on the next page

Table S1 – continued from previous page

Pre-treatments	Model	Soil property	R^2	MSE	RMSE	bias	RPD
SVG-1-2-11 + NR 434	plsr	Ca	-1.555	0.112	0.324	0.02	0.695
SVG-1-2-11 + NR 434	rf	Ca	-0.502	0.083	0.265	0.021	0.881
SVG-1-2-11 + NR 434	ann	K	0.048	0.02	0.133	0	1.057
SVG-1-2-11 + NR 434	cb	K	0.086	0.022	0.132	-0.008	1.113
SVG-1-2-11 + NR 434	plsr	K	-0.291	0.022	0.144	-0.001	0.951
SVG-1-2-11 + NR 434	rf	K	0.186	0.017	0.12	0.007	1.167
SVG-1-2-11 + NR 434	ann	Mg	-0.174	0.105	0.32	-0.014	0.947
SVG-1-2-11 + NR 434	cb	Mg	-0.258	0.114	0.33	-0.021	0.936
SVG-1-2-11 + NR 434	plsr	Mg	-0.083	0.094	0.305	0.012	0.989
SVG-1-2-11 + NR 434	rf	Mg	0.177	0.076	0.269	0.014	1.137
SVG-1-2-11 + NR 434	ann	Na	-2.559	0.005	0.05	0.003	0.827
SVG-1-2-11 + NR 434	cb	Na	-3.351	0.004	0.049	-0.002	0.755
SVG-1-2-11 + NR 434	plsr	Na	-6.623	0.005	0.063	0	0.51
SVG-1-2-11 + NR 434	rf	Na	-2.143	0.003	0.043	0.004	0.837
SVG-1-2-11 + NR 434	ann	P	-0.61	107.351	9.36	0.601	0.897
SVG-1-2-11 + NR 434	cb	P	-0.203	80.812	8.361	-1.931	0.932
SVG-1-2-11 + NR 434	plsr	P	-1.014	109.473	10.118	-0.289	0.763
SVG-1-2-11 + NR 434	rf	P	-0.007	68.8	7.623	0.483	1.042
SVG-1-2-11 + NR 434	ann	pH	-0.009	0.152	0.386	0.006	1.044
SVG-1-2-11 + NR 434	cb	pH	0.032	0.145	0.38	-0.014	1.049
SVG-1-2-11 + NR 434	plsr	pH	-0.064	0.157	0.393	0.005	1.027
SVG-1-2-11 + NR 434	rf	pH	0.352	0.098	0.312	-0.007	1.278
SVG-1-2-9	ann	Al	-0.039	2.059	1.413	0.08	1.038
SVG-1-2-9	cb	Al	0.145	1.694	1.282	-0.062	1.149
SVG-1-2-9	plsr	Al	0.095	1.684	1.28	-0.036	1.165
SVG-1-2-9	rf	Al	0.511	0.981	0.973	0.045	1.516
SVG-1-2-9	ann	TC	0.49	11.863	3.192	0.079	1.985
SVG-1-2-9	cb	TC	0.717	7.869	2.701	-0.087	2.148
SVG-1-2-9	plsr	TC	0.596	9.517	3.046	-0.107	1.845
SVG-1-2-9	rf	TC	0.833	4.84	2.147	0.014	2.604
SVG-1-2-9	ann	H	0.221	0.411	0.629	-0.012	1.208
SVG-1-2-9	cb	H	0.281	0.371	0.602	0.009	1.246
SVG-1-2-9	plsr	H	0.476	0.272	0.519	-0.014	1.426
SVG-1-2-9	rf	H	0.574	0.22	0.466	0.007	1.596
SVG-1-2-9	ann	N	0.599	0.041	0.186	-0.006	2.05
SVG-1-2-9	cb	N	0.694	0.034	0.179	-0.009	1.919
SVG-1-2-9	plsr	N	0.54	0.043	0.206	-0.007	1.635
SVG-1-2-9	rf	N	0.791	0.021	0.143	0.003	2.37
SVG-1-2-9	ann	Ca	-0.728	0.084	0.265	0	0.915
SVG-1-2-9	cb	Ca	-1.905	0.13	0.329	-0.013	0.771
SVG-1-2-9	plsr	Ca	-1.643	0.113	0.325	0.018	0.693
SVG-1-2-9	rf	Ca	-0.337	0.079	0.254	0.016	0.947
SVG-1-2-9	ann	K	-0.11	0.023	0.142	-0.005	1.027
SVG-1-2-9	cb	K	0.223	0.019	0.123	-0.017	1.19
SVG-1-2-9	plsr	K	-0.192	0.02	0.138	0	0.988
SVG-1-2-9	rf	K	0.26	0.016	0.116	0.006	1.225
SVG-1-2-9	ann	Mg	-0.275	0.122	0.332	0.009	0.949
SVG-1-2-9	cb	Mg	-0.289	0.116	0.334	-0.029	0.922
SVG-1-2-9	plsr	Mg	-0.224	0.108	0.325	0.012	0.936
SVG-1-2-9	rf	Mg	0.191	0.074	0.267	0.012	1.146
SVG-1-2-9	ann	Na	-0.54	0.003	0.039	0.001	0.903
SVG-1-2-9	cb	Na	-2.079	0.003	0.045	0	0.862
SVG-1-2-9	plsr	Na	-8.53	0.005	0.066	0.001	0.491
SVG-1-2-9	rf	Na	-1.422	0.003	0.041	0.004	0.845
SVG-1-2-9	ann	P	-0.217	84.147	8.449	0.102	0.94
SVG-1-2-9	cb	P	-0.163	78.572	8.223	-2.042	0.952
SVG-1-2-9	plsr	P	-1.217	116.174	10.524	-0.239	0.73
SVG-1-2-9	rf	P	-0.002	69.391	7.65	0.502	1.036
SVG-1-2-9	ann	pH	-0.125	0.169	0.409	-0.005	0.972
SVG-1-2-9	cb	pH	-0.058	0.16	0.398	-0.011	1.004
SVG-1-2-9	plsr	pH	-0.094	0.164	0.402	0	0.996
SVG-1-2-9	rf	pH	0.352	0.098	0.312	-0.006	1.277

The description of each pre-treatment is on Table 2.1; rf: random Forest; cb: Cubist; plsr: Partial Least Squares Regression; TC: Total Carbon; R^2: coefficient of determination; MSE: Mean Squared Error; RMSE: Root Mean Square Error; RPD: Ratio of Performance to Deviation. The coefficients units correspond to Table 2.2.

APPENDIX B — Acquired CHRIS PROBA images

Complementary information to the Chapter II.

Table S2: List of acquired CHRIS PROBA images.

Image code	Selected images
CHRIS_I3_170608_38AD_41	no
CHRIS_I3_170608_38AD_41.hdf	no
CHRIS_I3_170608_38AE_41.hdf	no
CHRIS_I3_170608_38AF_41.hdf	no
CHRIS_I3_170608_38B0_41.hdf	no
CHRIS_I3_170608_38B1_41.hdf	no
CHRIS_I3_170713_3A60_41	no
CHRIS_I3_170713_3A60_41.hdf	yes
CHRIS_I3_170713_3A61_41.hdf	yes
CHRIS_I3_170713_3A62_41.hdf	no
CHRIS_I3_170713_3A63_41.hdf	no
CHRIS_I3_170713_3A64_41.hdf	no
CHRIS_I3_180811_4CDF_41	no
CHRIS_I3_180811_4CDF_41.hdf	yes
CHRIS_I3_180811_4CE0_41.hdf	no
CHRIS_I3_180811_4CE1_41.hdf	no
CHRIS_I3_180811_4CE2_41.hdf	no

Image codification:
<Instrument>_<TargetCode>_<YYMMDD>_<ImageID>_
<Version>.<FileType>.