MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL RURAL DA AMAZÔNIA – UFRA
EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA - EMBRAPA

LUCIANA MARIA DE OLIVEIRA

IMPUTAÇÃO MÚLTIPLA E FUNÇÕES DE PEDOTRANSFERÊNCIA PARA ESTIMATIVA DA DENSIDADE DE SOLOS DA AMAZÔNIA ORIENTAL

BELÉM
2019

LUCIANA MARIA DE OLIVEIRA

IMPUTAÇÃO MÚLTIPLA E FUNÇÕES DE PEDOTRANSFERÊNCIA PARA ESTIMATIVA DA DENSIDADE DE SOLOS DA AMAZÔNIA ORIENTAL

Tese apresentada a Universidade Federal Rural da Amazônia, como parte das exigências do curso de Doutorado em Agronomia, para obtenção do título de Doutor.

Área de concentração: Agronomia

Orientadora: Dr[a]. Herdjania Veras de Lima

Coorientador: Dr. Eduardo Jorge Maklouf Carvalho-EMBRAPA

BELÉM

2019

Dados Internacionais de Catalogação na Publicação (CIP)
Bibliotecas da Universidade Federal Rural da Amazônia
Gerada automaticamente mediante os dados fornecidos pelo(a) autor(a)

O48 Oliveira, Luciana Maria de

IMPUTAÇÃO MÚLTIPLA E FUNÇÕES DE PEDOTRANSFERÊNCIA PARA ESTIMATIVA DA DENSIDADE DE SOLOS DA AMAZÔNIA ORIENTAL / Luciana Maria de Oliveira. - 2019.

72 f. : il.

Tese (Doutorado) - 0, Campus Universitário de Belém, Universidade Federal Rural da Amazõnia, Belém, 2019.
Orientador: Profa. Dra. Herdjania Veras de Lima

1. Dados faltantes. 2. Imputação multivariada por equações encadeadas. 3. Banco de dados de solos. 4. Regressão linear múltipla. I. Lima, Herdjania Veras de , *orient.* II. Título

CDD 631.43

LUCIANA MARIA DE OLIVEIRA

IMPUTAÇÃO MÚLTIPLA E FUNÇÕES DE PEDOTRANSFERÊNCIA PARA ESTIMATIVA DA DENSIDADE DE SOLOS DA AMAZÔNIA ORIENTAL

Tese apresentada à Universidade Federal Rural da Amazônia, como parte das exigências do Curso de Doutorado em Agronomia, área de concentração Agronomia, para obtenção do título de Doutor.

Orientadora: Profª. Dra. Herdjania Veras de Lima
Coorientador: Eduardo Jorge Maklouf Carvalho

Aprovado em 21 de agosto de 2019.

BANCA EXAMINADORA

Dra. Herdjania Veras de Lima - Orientadora
UNIVERSIDADE FEDERAL RURAL DA AMAZÔNIA – UFRA

Dra. Sueli Rodrigues – 1º Examinador
UNIVERSIDADE FEDERAL DO PIAUÍ - BOM JESUS– UFPI

Dr. Edson Marcos Leal Soares Ramos - 2º Examinador
UNIVERSIDADE FEDERAL DO PARÁ – UFPA

Dr. Antônio Vinicius Correa Barbosa - 3º Examinador
UNIVERSIDADE FEDERAL RURAL DA AMAZÔNIA – UFRA

Dra. Lorena Chagas Torres - 4º Examinador
UNIVERSIDADE FEDERAL RURAL DA AMAZÔNIA – UFRA

Dedico este trabalho, a minha mãe, Maria Alexandre da Silva (in memoriam), por ter sido modelo de coragem, perseverança, integridade e dedicação.

AGRADECIMENTOS

Primeiro de tudo, gostaria de agradecer a Deus por me guiar, iluminar e me dar tranquilidade para seguir em frente com os meus objetivos e não desistir com as dificuldades.

Ao meu esposo, Max Roberto de Souza Santos, por todo o amor, paciência, torcida e companheirismo em todos os momentos.

A Professora e Orientadora, Dra. Herdjania Veras de Lima, pela receptividade, confiança e por sua orientação.

Ao pesquisador e coorientador Dr. Eduardo Jorge Maklouf Carvalho, pela parceria que viabilizou este trabalho.

A Professora Dra. Sueli Rodrigues, pela partilha do saber, competência e pelas valiosas contribuições para o trabalho.

Ao Professor Dr. Edson Marcos Leal Soares Ramos, pelas relevantes sugestões.

Ao Professor Dr. Antônio Vinicius Correa Barbosa, meus agradecimentos por seus questionamentos e contribuições.

A Pesquisadora Lorena Torres, pelas preciosas contribuições nesta pesquisa.

Ao Professor Dr. Pedro Silvestre da Silva Campos, meus agradecimentos por seu apoio.

Ao Professor Dr. Rosemiro dos Santos Galate, pelo incentivo e apoio.

À minha família, pelas orações, torcida e confiança que sempre depositam em mim.

À Universidade Federal Rural da Amazônia e ao Programa de Pós-Graduação em Agronomia - PGAGRO pela acolhida e oportunidades oferecidas.

Ao Instituto Ciberespacial - ICIBE, representado pelo Diretor Prof. Dr. Pedro Silvestre da Silva Campos, pelo apoio e por ter concedido meu afastamento para que eu pudesse me dedicar ao doutorado.

A todos os professores que fizeram parte deste caminhar.

Aos meus amigos e colegas de trabalho: Aline, Daynara, Deciola e Peola, pela amizade e auxilio.

A todas as pessoas que contribuíram para a concretização desta tese, estimulando-me intelectualmente, emocionalmente e espiritualmente.

A todos,
O meu Eterno Obrigado.

LISTA DE ILUSTRAÇÕES

CONTEXTUALIZAÇÃO

IMPUTAÇÃO MÚLTIPLA PARA PREENCHIMENTO DE LACUNAS EM BANCO DE DADOS DE PROPRIEDADES FÍSICO HÍDRICAS DE SOLO

FUNÇÕES DE PEDOTRANSFERÊNCIA PARA ESTIMAR A DENSIDADE DO SOLO NA AMAZÔNIA ORIENTAL, BRASIL

LISTA DE TABELAS

IMPUTAÇÃO MÚLTIPLA PARA PREENCHIMENTO DE LACUNAS EM BANCO DE DADOS DE PROPRIEDADES FÍSICO HÍDRICAS DE SOLO

FUNÇÕES DE PEDOTRANSFERÊNCIA PARA ESTIMAR A DENSIDADE DO SOLO NA AMAZÔNIA ORIENTAL, BRASIL

LISTA DE ABREVIAÇÕES E SIGLAS

ACC	Análise de casos completos
AIC	Critério de informação de Akaike
CC	Capacidade de campo (-6 kPa)
COS	Carbono orgânico do solo
Dp	Densidade de partícula
Ds	Densidade do solo
Ds_{mi}	Densidade do solo medida
Ds_{pi}	Densidade do solo estimada
EMBRAPA	Empresa Brasileira de Pesquisa Agropecuária
EP	Erro padrão
EM	Erro médio
EQM	Erro médio quadrático
FMI	Fração de informações faltantes
IC (95%)	intervalo com 95% de confiança
Macro	Macroporosidade
Máx	Máximo
EM	Erro Médio
Micro	Microporosidade
MICE	*Multivariate Imputation by Chained Equations*
Mín	Mínimo
MO	Matéria orgânica
MSE	erro médio ao quadrado
RMSE	Raiz do quadrado do erro médio
PMP	Ponto de murcha permanente (-1500 kPa)
PT	Porosidade total
FPT	Função de pedotransferência
P-valor	Probabilidade de se observar um valor da estatística de teste maior ou igual ao encontrado
RML	Regressão linear múltipla

LISTA DE SÍMBOLOS

β	Coeficientes
λ	Proporção da variância total que é atribuível aos dados faltantes.
r	Coeficiente de correlação de Pearson
R^2	Coeficiente de determinação
R^2_{aj}	Coeficiente de determinação ajustado
$\hat{Y}$	Modelo estimado
* e **	Significativo a 5 e 1% de probabilidade, respectivamente
kPa	Quilopascal
p	Número de parâmetros do modelo
n	Número de amostras
d	Índice de Willmott
C	Índice de desempenho

RESUMO

A densidade do solo (Ds) é um atributo físico básico na ciência do solo, muito importante devido sua relação com diversas propriedades e processos do solo, tais como: propriedades hidráulicas, compactação do solo e erosão. É um atributo dinâmico que varia conforme o tipo de solo, manejo, cobertura vegetal, etc. Apesar da importância da Ds, é comum encontrar bancos de dados em todo o mundo que não tenham medições deste atributo para todos ou alguns registros, devido às medições serem trabalhosas, pois requer coleta de amostra de solo indeformada (difícil obtenção). Para solucionar esse problema, funções de pedotransferência são usadas para estimar atributos do solo a partir de dados disponíveis de levantamentos de solo. Outro problema muito frequente é a ocorrência de conjuntos de variáveis incompletas, devido a diversos fatores como: erro de digitação, não medição de determinado atributo em certo local (amostras provenientes de diversos locais). O objetivo geral deste trabalho foi avaliar a eficiência do método MICE (Multivariate Imputation by Chained Equations) para preencher banco de dados de atributos físico-hídricos dos solos com dados faltantes e desenvolver funções de pedotransferência para a estimativa da densidade do solo. Os dados utilizados foram compilados de diversas fontes, entre os anos de 1997 a 2014 para compor o banco de dados, que continham duas classes de solo (argissolos e latossolos) na profundidade de 0 a 60 cm, formado por 631 amostras e, onze variáveis, das quais cinco continham dados faltantes. Estas foram tratadas por meio da imputação de dados faltantes usando modelos de regressão linear. Com a imputação o tamanho do banco de dados e a distribuição geral dos dados foram preservados. A densidade do solo (Ds) foi obtida por meio de funções de pedotransferência (FPT), utilizando duas composições do banco de dados: com imputação e sem imputação de dados. As variáveis preditoras foram selecionadas pelo método “stepwise”. O melhor desempenho foi obtido para a FPT1 (sem imputação), que utilizou o teor de argila e a macroporosidade como covariáveis e, para a FPT4 (com imputação) que usou como covariáveis a matéria orgânica do solo e o silte. A FPT4 é a mais indicada por utilizar covariáveis mais facilmente encontrada em banco de dados de solos. Das FPTs disponíveis na literatua a de Tomasella e Hodnett (1998) foi que apresentou melhor performance. A estimativa de atributos do solo utilizando banco de dados com imputação tem potencial para solucionar problemas de falta de dados de solos da região amazônica.

Palavras-chave: Dados faltantes. Imputação multivariada por equações encadeadas. Banco de dados de solo. Regressão linear múltipla.

ABSTRACT

Soil density (Ds) is a basic physical attribute in soil science, very important because of its relationship to various soil properties and processes such as hydraulic properties, soil compaction and erosion. It is a dynamic attribute that varies according to soil type, management, vegetation cover, etc. Despite the importance of Ds, it is common to find databases around the world that do not have measurements of this attribute for all or some records, because the measurements are laborious because it requires undisturbed soil sample collection (difficult to obtain). To solve this problem, pedotransfer functions are used to estimate soil attributes from available ground survey data. Another very common problem is the occurrence of incomplete sets of variables, due to several factors such as: typo, not measuring a certain attribute in a certain place (samples from different places). The general objective of this work was to evaluate the efficiency of the Multivariate Imputation by Chained Equations (MICE) method to fill the soil physical-water attribute database with missing data and to develop pedotransfer functions for soil density estimation. The data used were compiled from several sources, from 1997 to 2014 to compose the database, which contained two soil classes (argisols and latosols) at a depth of 0 to 60 cm, consisting of 631 samples and eleven variables. , of which five contained missing data. These were treated by imputing missing data using linear regression models. With imputation the size of the database and the overall distribution of the data were preserved. Soil density (Ds) was obtained by pedotransfer functions (FPT) using two database compositions: with imputation and without imputation of data. Predictor variables were selected by the stepwise method. The best performance was obtained for FPT1 (without imputation), which used the clay content and macroporosity as covariates, and for FPT4 (with imputation) that used as soil organic matter and silt as covariates. FPT4 is best suited for using covariates most easily found in a soil database. Of the FPTs available in the literature by Tomasella and Hodnett (1998) presented the best performance. Estimation of soil attributes using imputation database has the potential to solve problems of lack of soil data in the Amazon region.

Keywords: Missing data. Multivariate imputation by chained equations. Soil database. Multiple linear regression.

SUMÁRIO

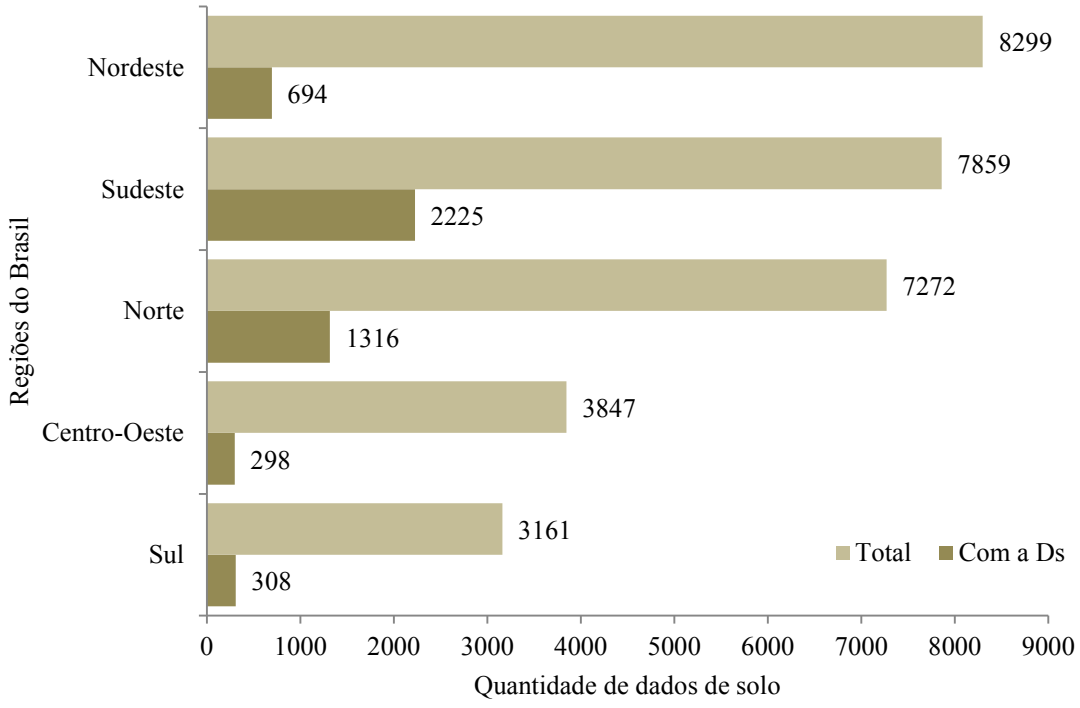
1 CONTEXTUALIZAÇÃO

A densidade do solo (Ds) é um atributo físico básico na ciência do solo, muito importante devido sua relação com diversas propriedades e processos do solo, tais como: propriedades hidráulicas, compactação do solo e erosão (CHAUDHARI et al., 2013). É um atributo dinâmico que varia conforme o tipo de solo, manejo, cobertura vegetal, etc. (CARVALHO JUNIOR et al., 2016).

Apesar da importância da Ds, é comum encontrar bancos de dados em todo o mundo que não tenham medições deste atributo para todos ou alguns registros, devido às medições serem trabalhosas (requer coleta de amostra de solo indeformada), demoradas e caras (SEQUEIRA et al., 2014).

No Brasil, a maior base de dados de solos é o Sistema de informação de solos brasileiros disponibilizado pela Embrapa. Embora haja uma considerável quantidade de dados, ainda há carência de informações em relação aos valores de Ds. Dentre as regiões brasileiras, a região norte ocupa o terceiro lugar em relação a quantidade total de dados de solos. No entanto, ocupa segundo quando se trata da disponibilidade de dados de Ds (Figura 1).

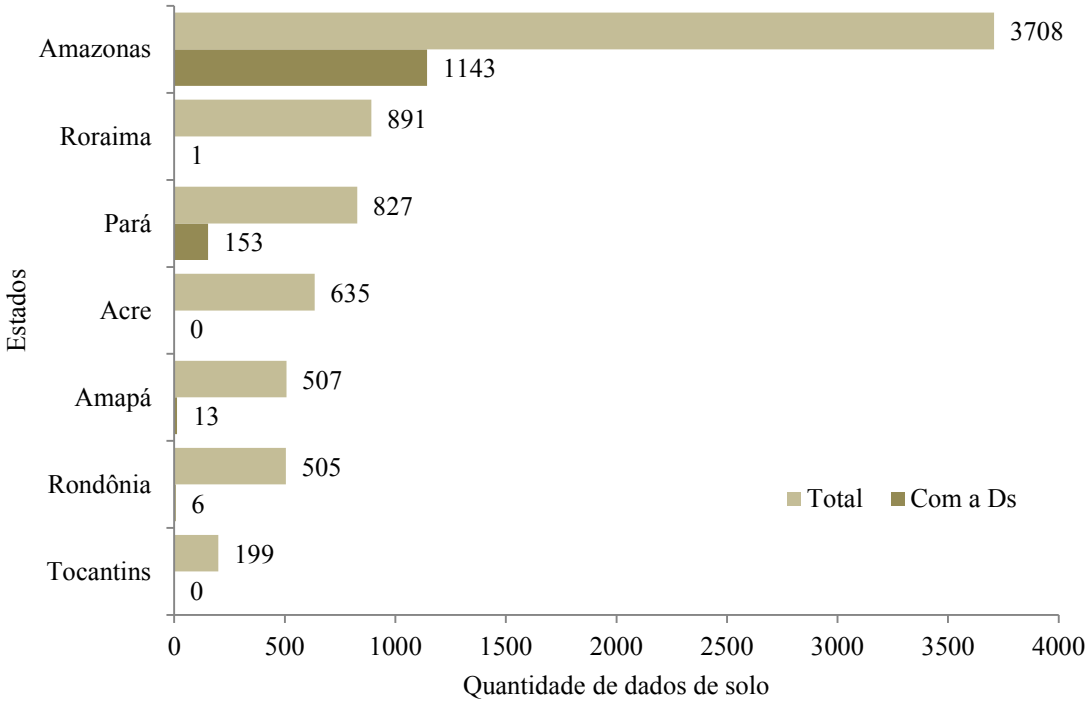
Figura 1 - Quantidade dos dados de solos por regiões do Brasil, obtidos a partir do Sistema de informações de solos brasileiros da Embrapa (Barra mais escura = com Ds e barra mais clara = total de dados).



Fonte: Dados SISolos - Embrapa, 2019.

Na região Norte, o Estado do Pará contém a segunda maior quantidade de dados Ds (Figura 2).

Figura 2 - Quantidade de dados de solos da região norte do Brasil, por Estado, a partir do Sistema de informações de solos brasileiros da Embrapa (Barra mais escura = com Ds e barra mais clara = total de dados).



Fonte: Dados SISolos - Embrapa, 2019.

Tendo em vista a necessidade de informações sobre a Ds, as funções de pedotransferência (FPT) apresentam-se como uma alternativa para estimar este atributo do solo a partir de dados mais facilmente disponíveis em estudos de levantamento de solos ou mais facilmente medidos (VAN LOOY et al., 2017). Atualmente, na física do solo as FPTs são aplicadas principalmente para estimar o teor de água no solo em diferentes tensões, (KARUP et al., 2016; AULER, PIRES e PINEDA, 2017; CONTRERAS e BONILLA, 2018), a densidade do solo (CHEN et al., 2018) e a condutividade hidráulica do solo (GHANBARIAN, TASLIMITEHRANI e PACHEPSKY, 2017). Várias funções de pedotransferência (FPTs) foram desenvolvidas para prever a Ds em diversas regiões do mundo (LU et al., 2019) como, França (CHEN et al., 2018;), Grécia (SEVASTAS et al., 2018), China (QIAO et al., 2018), em toda a República da Irlanda (PREMROV, CUMMINS e BYRNE, 2018), entre outras. Bem como em diversas regiões Brasil (BEUTLER et al. , 2017) , Rio de Janeiro (Carvalho Junior et al., 2016), Minas Gerais (Souza et al., 2016), Amazônia

Central (Barros et al., 2015). No entanto, não há registros de FPTs para estimar a Ds para o Estado do Pará.

A precisão e a confiabilidade das FPTs são altamente dependentes das características dos dados (escala, variáveis preditoras, tamanho da amostra, heterogeneidade, entre outros) e de técnicas usadas em seu desenvolvimento (KHODAVERDILOO et al., 2018; TÓTH et al., 2015). As bases de dados utilizadas no desenvolvimento de FPTs podem não refletir as características dos solos de outras regiões, e como resultado, podem apresentar baixa precisão (valores medidos estão longe uns dos outros) e/ou baixa exatidão (valores medidos estão longe do valor verdadeiro, ou de referência) quando aplicadas em regiões para as quais não foram desenvolvidas (SCHAAP e LEIJ, 1998). Dessa forma, a escolha da FPT adequada ao solo baseada nas características dos solos de cada região e da dependência espacial de seus atributos são fundamentais para predição com o menor desvio nas previsões (SILVA e ARMINDO, 2016). Além disso, algumas FPTs foram desenvolvidas com base em um número limitado de dados (KHODAVERDILOO et al., 2018).

Uma limitação muito frequente, porém, ainda pouco explorada na área da Ciência do solo é a ocorrência de conjunto de dados com amostras incompletas isto é, amostras que contém valores ausentes, devido a diversos fatores como: erros de digitação, não medição de um atributo em um determinado local (amostras proveniente de diversos locais), etc. Tal fato pode impossibilitar o uso de algumas técnicas estatísticas projetadas para análise de dados completos, como a análise multivariada (análise de regressão linear múltipla), a qual, exige completude nas matrizes dos dados, uma vez que, valores ausentes podem causar viés, além de reduzir a eficiência dos modelos (MADLEY-DOWD et al., 2019).

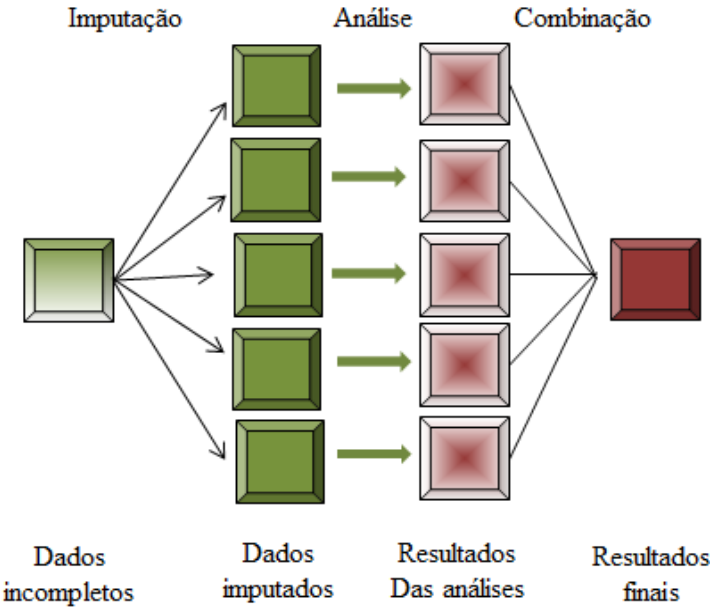
A remoção de amostras (*listwise deletion*) com dados faltantes tem sido a prática mais comum (e o padrão na maioria dos pacotes estatísticos), restringindo-se assim, à análise dos casos completos - ACC (VAN GINKEL et al., 2019). Poucos autores relatam a ocorrência de dados faltantes nos bancos de dados que originaram as FPTs para solos do Brasil. Pode-se destacar o estudo de Benites et al. (2007), onde o número de observações utilizadas no desenvolvimentos de FPTs para estimar a Ds no Brasil variou entre 1342 e 1539, devido à falta de alguns dados. Medeiros et al. (2014) relataram a retirada de 17,3% de amostras com dados faltantes para a criação de FPTs para predizer a curva de retenção de água no solo no estado do Pará.

Uma maneira de remediar o desperdício de dados é “substituir” os valores faltantes por valores plausíveis ("valores imputados"), técnica conhecida como imputação de dados. A

principio, a imputação era única e substituía os dados faltantes pela média ou pela mediana da variável, por interpolação ou até por regressão linear. Em seguida, surgiram as técnicas de imputação múltipla para corrigir incertezas associadas à imputação única (VAN BUUREN e GROOTHUIS-OUDSHOORN, 2011).

O método de imputação múltipla se resume em três etapas principais: imputação, análise e agrupamento (Figura 3). Para considerar a incerteza sobre os valores imputados, vários conjuntos de dados completos são criados. Estes são analisados separadamente usando método estatístico padrão e os múltiplos conjuntos de resultados são combinados usando as "regras de Rubin" (RUBIN, 1987).

Figura 3 - Principais etapas da imputação múltipla.



Fonte: Adaptado de van Buuren e Oudshoorn, 1999.

O uso da imputação mútipla (IM) é atualmente uma das ferramentas mais eficientes no preenchimento de dados faltantes. Embora, do ponto de vista teórico, a IM seja considerada o método ideal, muitos pesquisadores ainda são resistentes ao uso do método, devido a falta de familiaridade com a técnica (VAN GINKEL et al., 2019). O mais comum é o uso da exclusão listwise, que consiste em remover todos os dados para cada caso que tenha um ou mais valores faltantes. Essa análise é chamada de Análise Completa de Casos (ACC).

Poucos estudos em ciências agrárias tratam o problema de dados faltantes. Clifford et al. (2014) utilizaram um método de imputação não paramétrica para “preencher” os dados faltantes em um banco de dados de atributos químicos e físicos do solo em aproximadamente 9500 locais na Austrália. Eles indicaram que o processo de imputação proposto pode ser estendido a outros cenários na ciência do solo. No Brasil, Arciniegas-Alarcón et al. (2014)

propuseram um novo algoritmo de imputação múltipla de livre distribuição e sem restrições quanto ao padrão ou mecanismo de ausência dos dados para contornar a ocorrência de dados faltantes em experimentos genótipo x ambiente (GxE) e afirmaram que o método pode ser utilizado em qualquer conjunto de dados que se adeque ao arranjo matricial.

Dentre os métodos de imputação múltipla, a imputação multivariada por equações encadeadas (MICE) tem sido o mais difundido. Este gera estimativa usando a comparação média preditiva, regressão linear bayesiana, regressão logística e outras (VAN BUUREN e GROOTHUIS-OUDSHOORN, 2011). A partir de amostragem aleatória, o MICE executa sequencialmente imputações univariadas até a convergência. Cada interação é um amostrador de Gibbs que é extraído da distribuição condicional nos valores imputados (BERTSIMAS, PAWLOWSKI e ZHUO, 2018). Não há relatos de estudos que utilizam a IM (MICE) na ciência do solo.

Uma das principais restrições ao desenvolvimento de uma FPT é a indisponibilidade de banco de dados de solos extensos (BOTULA et al., 2014), onde o número de amostras seja grande o suficiente ($n \geq 500$) para desenvolver FPTs precisas e confiáveis, uma vez que, FPTs desenvolvidas a partir de bancos de dados maiores são mais robustas do que as FPTs derivadas de bancos de dados regionais menores (KHODAVERDILOO et al., 2018).

Face à carência de informações disponíveis sobre a Ds para região Amazônica, devido a sua dimensão geográfica e as condições de acesso a essas áreas, ferramentas estatísticas como métodos de imputação de dados e funções de pedotransferência são importantes. Assim, as hipóteses testadas neste estudo foram: (i) é mais eficiente realizar a imputação múltipla dos dados do que restringir-se à análise dos casos completos (ACC); (ii) o desempenho das FPTs para estimar a Ds desenvolvidas neste estudo é superior a outras disponíveis na literatura para a região da Amazônia Oriental.

A presente tese tem como objetivo principal, estimar valores faltantes utilizando o método MICE em banco de dados de atributos físico-hídricos dos solos da região da Amazônia Oriental e desenvolver funções de pedotransferência por meio de regressão linear múltipla capaz de estimar de forma eficiente a densidade do solo. Para atender os objetivos propostos, a tese foi subdividida em dois capítulos. O primeiro capítulo apresenta o método de imputação múltipla (MICE) para "preenchimento" dos dados faltantes em banco de dados de propriedades físico-hídricas de solo. O método utilizado foi a imputação multivariada por equações encadeadas (MICE - *Multivariate Imputation by Chained Equations*), um dos

muitos algoritmos que realizam imputações múltiplas (IM) com base na cadeia de Markov Monte Carlo (MCMC).

O segundo capítulo apresenta o desenvolvimento e avaliação de FPTs para estimar a densidade do solo (Ds) na Amazônia oriental; e compara o desempenho das PTFs com modelos disponíveis na literatura.

Em seguida são apresentadas as conclusões mais relevantes acerca dos principais aspectos abordados na pesquisa, assim como algumas recomendações para o direcionamento de investigações e estudos futuros.

No Apêndice está apresentado o algoritmo de imputação múltipla dos dados, desenvolvidos pelo programa computacional R utilizando o método MICE - Multivariate Imputation by Chained Equations (VAN BUUREN e GROOTHUIS-OUDSHOORN, 2011). E os gráficos das análises dos resíduos das funções de pedotransferência desenvolvidas.

REFERÊNCIAS

ARCINIEGAS-ALARCÓN, S.; DIAS, C.T.S.; GARCÍA-PEÑA, M. Distribution free multiple imputation in incomplete two-way tables. **Pesquisa Agropecuária Brasileira**, Brasília, v.49, n.9, p.683-691, set. 2014.

AULER, A.C.; PIRES, L.F.; PINEDA, M.C. Influence of physical attributes and pedotransfer function for predicting water retention in management systems, Campina Grande, PB. **Revista Brasileira de Engenharia Agrícola e Ambiental**, v.21, p.746-51, sep. 2017.

BENITES, V. M.; MACHADO, P. L.; FIDALGO, E. C.; COELHO, M. R.; MADARI, B. E. Pedotransfer functions for estimating soil bulk density from existing soil survey reports in Brazil. **Geoderma**, v. 139, p. 90-97, 2007.

BERTSIMAS, D.; PAWLOWSKI, C.; ZHUO, Y.D. From predictive methods to missing data imputation: An optimization approach. **Journal of Machine Learning Research**, v.18, p. 1-39, apr.2018. Disponível em: <http://jmlr.org/papers/v18/17-073.html>. Acesso em: 02 jul. 2019.

BOTULA, Y.D.; RANST, E.V.; CORNELIS, W.M. Pedotransfer functions to predict water retention for soils of the humid tropics: a review. **Revista Brasileira de Ciência do Solo**, v.38, p.679-698, 2014.

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

CHAUDHARI, P.R.; AHIRE, D. V.; AHIRE, V. D.; CHKRAVARTY, M.; MAITY, S. Soil Bulk Density as related to Soil Texture, Organic Matter Content and available total Nutrients of Coimbatore Soil. **International Journal of Scientific and Research Publications**, v. 3, p.1-8, feb. 2013.

CHEN, S.; RICHER-DE-FORGES, A. C.; SABY, N. P. A.; MARTIN, M. P.; C. WALTER, C.; ARROUAYS, D. Building a pedotransfer function for soil bulk density on regional dataset and testing its validity over a large area. **Geoderma**,
vol. 312, p. 52–63, feb. 2018.

CLIFFORD D, DOBBIE MJ, SEARLE R. Non-parametric imputation of properties for soil profiles with sparse observations. **Geoderma**. Australia, v.232–234. p.10–18, apr, 2014.

CONTRERAS, C.P.; BONILLA, C.A. A comprehensive evaluation of pedotransfer functions for predicting soil water content in environmental modeling and ecosystem management. **Science of the Total Environment**, v. 644, p. 1580–1590, dec. 2018.

EMBRAPA. SISolos - Sistema de Informação de Solos Brasileiros. 2014. Disponível em: <http://www.sisolos.cnptia.embrapa.br>. Acesso em: 02 jul. 2019.

GHANBARIAN, B.; TASLIMITEHRANI, V.; PACHEPSKY, Y. A. Accuracy of sample dimension-dependent pedotransfer functions in estimation of soil saturated hydraulic conductivity. **Catena**, v. 149, p. 374–380, feb. 2017.

KARUP, D., MOLDRUP, P., TULLER, M. ARTHUR, E. AND DE JONGE, L.W. Prediction of the soil water retention curve for structured soil from saturation to oven-dryness. **European Journal of Soil Science,** v.68, p.57–65, dec. 2016.

KHODAVERDILOO, H.; MOMTAZ, H.; LIAO, K. Performance of Soil Cation Exchange Capacity Pedotransfer Function as Affected by the Inputs and Database Size. **Clean – Soil, Air, Water**, v.46, p. 1-8, mar. 2018.

LU, Y.; SI,B.; LI, H.; BISWAS, A. Elucidating controls of the variability of deep soil bulk density. **Geoderma**. Canada, v. 348,p. 146–15, apr. 2019.

MADLEY-DOWD, P.; HUGHES, R.; TILLING, K.; HERON, J. The proportion of missing data should not be used to guide decisions on multiple imputation. **Journal of Clinical Epidemiology**, v. 110, p. 63-73, mar. 2019.

MEDEIROS, J.C., COOPER, M., DALLA ROSA, J., GRIMALDI, M., COQUET, Y. Assessment of pedotransfer functions for estimating soil water retention curves for the amazon region. **Revista Brasileira de Ciência do Solo**, Viçosa, v. 38, p. –730-743, mar. 2014.

PREMROV, A.; CUMMINS, T.; BYRNE, K. Bulk-density modelling using optimal power-transformation of measured physical and chemical soil parameters. **Geoderma**, v.314, p.205-220, mar. 2018.

QIAO, J.; ZHU, Y.; JIA, X.; HUANG, L.; SHAO, M. Development of pedotransfer functions for predicting the bulk density in the critical zone on the Loess Plateau, China. **Journal of Soils and Sediments**, may. 2018.

RUBIN, D.B. **Multiple Imputation for Nonresponse in Surveys**. 1. ed. New York: John Wiley & Sons, 1987. 253 p.

SEQUEIRA, C.; WILLS, S.; SEYBOLD, C.; WEST, L. Predicting soil bulk density for incomplete databases. **Geoderma**, v.213, p.64–73, 2014.

SEVASTAS, S.; GASPARATOS, D.; BOTSIS, D.; SIARKOS, I.; DIAMANTARAS, K.I.; BILAS, G. Predicting bulk density using pedotransfer functions for soils in the Upper Anthemountas basin, Greece. **Geoderma Regional**, v.14, sep. 2018.

SILVA, A.C.; ARMINDO, R. A. Importância das Funções de Pedotransferência no estudo das propriedades e funções hidráulicas dos solos do Brasil. **Multi-Science Journal**, Goiania, v. 1, n. 5, p. 31 – 37, ago. 2016. Disponível em: < https://www.ifgoiano.edu.br/periodicos/index.php/multiscience/article/view/200 >. Acesso em: 02 jul. 2019.

SCHAAP, M. G.; LEIJ, F. J. Database-related accuracy and uncertainty of pedotransfer functions. **Soil Science**, v. 163(10), p.765–779, oct. 1998. Disponível em: <https://www.ars.usda.gov/arsuserfiles/20360500/pdf_pubs/P1696.pdf>. Acesso em: 02 jul. 2019.

TÓTH, B., WEYNANTS, M., NEMES, A., MAKÓ, A., BILAS, G., & TÓTH, G. New generation of hydraulic pedotransfer functions for Europe. **European Journal of Soil Science**, v.66(1), p.226–238, jan. 2015.

VAN BUUREN, S.; OUDSHOORN, K. Flexible multivariate imputation by MICE. **TNO Prevention and Health**, oct. 1999.

VAN BUUREN, S.; GROOTHUIS-OUDSHOORN, K. Mice: Multivariate Imputation by Chained. **Journal of Statistical Software**, v.45, dec. 2011.

VAN GINKEL, J. R.; LINTING, M.; RIPPE R. C. A. ; VAN DER VOORT, A. Rebutting existing misconceptions about multiple imputation as a method for handling missing data. **Journal of Personality Assessment**, online first. jan. 2019.

VAN LOOY, K., BOUMA, J., HERBST, M., KOESTEL, J., MINASNY, B., MISHRA, U., MONTZKA, C., NEMES, A., PACHEPSKY, Y., PADARIAN, J.: Pedotransfer functions in Earth system science: challenges and perspectives, **Reviews of Geophysics**., v. 55, p. 1199–1256, 2017.

2 IMPUTAÇÃO MÚLTIPLA PARA O PREENCHIMENTO DE DADOS FALTANTES EM BANCO DE DADOS DE PROPRIEDADES FÍSICO-HÍDRICAS DO SOLO

RESUMO

Valores faltantes em banco de dados é um problema comum e quase inevitável, entretanto, como os trabalhos lidam com ele raramente são mencionados. A imputação múltipla (IM) é um método eficaz para estimativas estatísticas de valores faltantes a partir de dados incompletos. O objetivo deste estudo foi avaliar a eficiência da IM utilizando o algoritmo MICE (Imputação Multivariada por Equações Encadeadas) para preencher dados ausentes em um banco de dados de propriedades físico-hídricas do solo e, mostrar que é viável realizar a imputação em vez da análise de casos completos (ACC). Análise preliminar do banco de dados foi realizada para verificar a adequação do algoritmo proposto. A imputação dos dados faltantes de cada variável foi ajustada usando modelos de regressão linear. Variáveis com dados faltante entram no modelo como variável dependente e as outras como covariáveis. As análises foram realizadas comparando os valores das estimativas, seus erros padrão e os intervalos de confiança de 95%. Concluiu-se que a IM apresentou melhor desempenho que a ACC, pois, embora a comparação estatística dos dois métodos seja semelhante, a IM mantém o tamanho do banco de dados e preserva a distribuição geral. Com este estudo, pretende-se verificar se o algoritmo MICE (*Multivariate Imputation by Chained Equations*) é uma boa alternativa para sanar a ausência de dados em bancos de dados de propriedades físico-hídricas de solos do Estado do Pará, e assim ajudar mais pesquisadores da ciência do solo a implementar a IM, em vez de abordagens de exclusão de dados como a Análise de casos completos (ACC), a fim de melhorar a precisão das análise estatísticas. Nosso estudo confirmou que a IM é aplicável a dados faltantes em banco de dados de propriedades do solo.

Palavras-chave: Imputação múltipla por equações encadeadas. Dados faltantes. Banco de dados de solo.

ABSTRACT

Missing database values is a common and almost inevitable problem, however, as jobs deal with it are rarely mentioned. Multiple imputation (MI) is an effective method for statistical estimation of missing values from incomplete data. The aim of this study was to evaluate the efficiency of MI using the MICE (Multivariate Chain Equation Imputation) algorithm to fill in missing data in a soil physical-water properties database and to show that it is feasible to perform imputation instead of analysis. Complete Case Studies (CCS). Preliminary analysis of the database was performed to verify the adequacy of the proposed algorithm. The imputation of missing data for each variable was adjusted using linear regression models. Missing data variables enter the model as the dependent variable and the others as covariates. Analyzes were performed comparing the estimate values, their standard errors and the 95% confidence intervals. It was concluded that MI performed better than CCS because, although the statistical comparison of the two methods is similar, MI maintains the size of the database and preserves the overall distribution. The aim of this study is to verify if the Multivariate Imputation by Chained Equations (MICE) algorithm is a good alternative to remedy the lack of data in the Pará State's soil physical-water properties databases, thus helping more researchers. soil science to implement MI rather than data exclusion approaches such as Full Case Analysis (CCS) to improve the accuracy of statistical analysis. Our study confirmed that MI is applicable to missing data in soil properties database.

Keywords: Multiple imputation by chained equations. Missing data. Soil database.

2.1 Introdução

Dados faltantes em pesquisas científicas são comuns e ocorrem por diversos motivos como, medidas não coletadas e erros e/ou imprecisão nas análises, gerando assim, bancos de dados incompletos, o que pode se tornar um obstáculo para análises estatísticas (AUDIGIER; HUSSON; JOSSE, 2015). No entanto, os problemas encontrados e as soluções implementadas são raramente mencionados na maioria das publicações. Isso pode ser devido à falta de importância dada ao problema (por ex. redução da amostra) ou por falta de conhecimento das soluções implementadas (muitas vezes de forma automática) pelos softwares estatísticos (FIGUEREDO *et al*., 2000).

Por exemplo, para análise de regressão múltipla, o procedimento padrão nos softwares estatístico, quando faltam dados, é a exclusão listwise, que consiste em remover todos os dados para cada caso que tenha um ou mais valores omissos. Essa análise é chamada de Análise Completa de Casos (CCA). Esse procedimento pode reduzir consideravelmente o banco de dados disponível e, assim, induzir a altos desvios preditivos na estimação dos parâmetros, contestando a validade das conclusões (PAES; POLETO, 2013).

O grau do problema é ainda mais significativo quando as análises multivariadas são implementadas, uma vez que essas análises exigem dados completos para todas as variáveis. (FIGUEREDO *et al*., 2000). Na ciência do solo, um exemplo é a estimativa de funções de pedotransferência que, a partir de amplos bancos de dados de solos, utilizam propriedades de fácil determinação como textura e densidade do solo para predizer outras mais complexas como aquelas relacionadas à capacidade de retenção de água no solo (SILVA; ARMINDO, 2016). Em geral, os dados usados para determinar funções de pedotransferência vêm de vários locais, portanto, dados ausentes são comuns.

Em alguns casos, uma opção para lidar com um banco de dados incompleto é preencher os valores ausentes usando métodos simples, como média, mediana, interpolação e regressão linear. Esses métodos são denominados imputação única (RUBIN, 1976). No entanto, a imputação única é limitada porque não leva em conta a incerteza associada à previsão de valores faltantes com base no valor observado (BUUREN, 2018).

Atualmente, procedimentos e softwares estatísticos modernos permitem um recurso mais eficaz para preencher essas lacunas. Um desses métodos é a Imputação Múltipla (IM), que considera a variabilidade entre as imputações, gerando conjuntos completos de dados, preenchendo os valores omissos por meio de modelos de imputação, geralmente mais precisos do que aqueles fornecidos pelos métodos de imputação única (LITTLE; RUBIN, 2015).

Embora a técnica de IM tenha sido utilizada em diversas áreas (CARVALHO et al., 2017; PEDERSEN et al., 2017; POYATOS et al., 2018; SQUILLANTE et al., 2018), na Ciência do Solo, ainda é pouco explorada (CLIFFORD; DOBBIE; SEARLE, 2014; SHAO; MENG; SUN, 2017).

Ao selecionar o método IM, recomenda-se que diferentes metodologias sejam exploradas de acordo com as características dos dados (KIM et al., 2015). A Imputação Multivariada por Equações Encadeadas (MICE) é um dos muitos algoritmos que realizam múltiplas imputações IM) com base na Cadeia de Markov de Monte Carlo (MCMC) (CARVALHO et al., 2017). Aplicações de MICE têm sido usadas em várias áreas, mas na Ciência do Solo esta abordagem ainda não foi usada.

O objetivo deste estudo é avaliar a eficiência do método de IM utilizando o algoritmo MICE (*Multivariate Imputation by Chained Equations*) para o preencher os valores faltantes em um banco de dados de propriedades físico-hídricas do solo e demonstrar que é mais viável realizar imputação do que restrita á Análise de casos completos (ACC).

2.2 Material e Métodos

2.2.1 Banco de dados do Solo

O banco de dados de solos (SDB) utilizado no estudo é proveniente de 24 municípios do estado do Pará, norte do Brasil. O SDB é composto por 631 amostras de duas classes de solos (Latossolos e Argissolos - SANTOS et al., 2013) amostradas nas profundidades de 0 a 60 cm entre 1997 e 2014. Os dados foram compilados de diversas fontes (trabalhos científicos, dissertações, teses, Boletins de Pesquisa da Embrapa e levantamentos de dados do solo realizados pela Embrapa Amazônia Oriental). Embora os SDB incluam variáveis quantitativas e qualitativas, apenas as seguintes variáveis quantitativas foram consideradas para este estudo: os teores de areia, silte e argila, determinados pelo método da pipeta; teor de matéria orgânica (MO), estimado pelo método Walkley-Black; densidade do solo (Ds) pelo método do núcleo; densidade de partículas (Dp) pelo método do picnômetro; porosidade total do solo (PT) pelo método de saturação; macroporosidade do solo (Macro); microporosidade do solo (Micro); teor de água no solo na capacidade de campo (CC) e ponto de murcha permanente (PMP), considerado como a umidade do solo equilibrada em potenciais de água de -6kPa e -1500kPa, respectivamente. Os dois últimos foram determinados no extrator da placa de pressão. Todas essas metodologias são descritas em Claessen et al. (1997).

2.2.2 Análise preliminar

Antes do processo de imputação, três análises preliminares dos dados faltantes foram realizadas para confirmar o padrão, o mecanismo e a proporção de desaparecimentos. As análises foram:

Padrão - o padrão de dados perdidos pode ser univariado (apenas uma variável contém dados perdidos) ou multivariada (mais de uma variável contém dados perdidos) (SONG; SHEPPERD, 2007). O padrão multivariado pode ocorrer como monótono ou arbitrário (RUBIN, 1987).

Se o padrão de dados perdidos for univariado, o método de imputação única (IU) é recomendado, enquanto o procedimento de imputação múltipla (IM) é recomendado para o padrão multivariado. Neste último, quando ocorre o padrão monótono, os métodos mais indicados são a Regressão Linear Bayesiana (BLR) e a Média Preditiva de Correspondência (PMM), enquanto que para o padrão arbitrário o método apropriado é a Cadeia de Markov de Monte Carlo (MCMC).

Mecanismo - os mecanismos de dados perdidos representam a relação estatística entre as observações (variáveis) e a probabilidade de dados perdidos e são classificados em três categorias (RUBIN, 1987): (i) Falta completamente aleatória (MCAR), quando a probabilidade dos dados em falta não dependem dos dados observados nem dos não observados; (ii) Falta aleatoriamente (MAR) quando a probabilidade de falta de dados depende em certa medida dos dados observados; e (iii) Não faltando aleatoriamente (NMAR), quando a probabilidade de dados perdidos depende dos valores de dados em falta.

Na prática, os dados perdidos quase nunca são MCAR, mas de alguma forma entre MAR e MNAR (GRAHAM, 2009). No entanto, os mecanismos MAR e NMAR não são identificados por testes. O mecanismo MCAR é testado pelo teste de Little (1988) e, quanto menor o valor de p ($p < 0,05$), mais forte é a evidência de que os dados não são MCAR.

Proporção - a proporção de dados perdidos foi verificada por meio do gráfico de barras. Se a proporção for $\leq 5\%$, o método de imputação única (IU) pode ser usado ou a análise completa do caso (ACC) pode ser considerada. Se a proporção for 5-15%, ainda é possível usar o método SI, no entanto, o método de imputação múltipla (IM) é recomendado. Quando a proporção de dados perdidos é $\geq 15\%$, o procedimento apropriado é o MI (HARRELL, 2016).

2.2.3 Imputação Múltipla

Verificadas as condições mostradas na subseção 2.2.2, o método escolhido foi a imputação múltipla por equações em cadeia (MICE), uma vez que mais de uma variável possui dados faltantes, nenhum padrão definido (padrão multivariado e arbitrário) foi observado e o mecanismo ausente é MAR.

O algoritmo MICE foi executado para o conjunto de variáveis (x) descritas na subseção 2.2.1, algumas ou todas com valores ausentes. O método consiste em executar uma série de modelos de regressão, onde cada variável dependente com dados perdidos é modelada em relação às outras variáveis do banco de dados (especificação totalmente condicional - FCS). Por meio de modelos de regressão linear ($\hat{y} = \beta_0 + \beta_1 x + ... + \beta_n x$), onde $\hat{y}$ é a variável a ser imputada.

O procedimento MICE pode ser dividido em três etapas principais: imputação, análise e combinação, descritas resumidamente abaixo:

Imputação - Geração de conjunto de dados completo. Os MICE geram uma série de estimativas onde cada variável, por sua vez, é regredida nas outras variáveis, isto é, percorre as variáveis que predizem cada variável dependendo das outras. O MICE é executado por meio de um processo iterativo: na primeira iteração, o modelo de imputação para a variável com o menor número de valores ausentes é estimado usando apenas dados completos. Em seguida, a variável com o segundo menor valor ausente é imputada usando os dados completos e os valores imputados da última iteração. Após cada variável ter passado por esse processo, o ciclo é repetido usando os dados da última iteração. Dez iterações foram realizadas onde os valores imputados após a 10ª e última iteração constituem um conjunto de dados imputado (STUART et al., 2009). Aqui, cinco versões de conjuntos de dados (m = 5) foram geradas, uma vez que, de acordo com Schafer e Olsen (1998), m de 3 a 5 é suficiente para obter estimativas precisas para a maioria das aplicações.

Análise - Separadamente, as cinco versões do conjunto de dados foram analisadas pelos métodos tradicionais de análise estatística (estimativas de parâmetros, erros padrão e intervalos de confiança de 95%).

Combinação - O último passo do MICE foi a combinação dos resultados das estimativas dos conjuntos de dados completos, usando o método de Rubin (1987). Cinco conjuntos diferentes de estimativas de ponto e variância para um parâmetro Q foram estimados. Seja Q_j e U_j as estimativas de ponto e variância do conjunto de dados imputado, j = 1, 2, ..., m. Então, a

estimativa pontual para Q de múltiplas imputações é a média das estimativas de dados completos:

$$\bar{Q} = \frac{1}{m}\sum_{j=1}^{m} Q_j \quad (1)$$

Seja $\bar{U}$ a variância dentro da imputação, que é a média das m estimativas completas de dados:

$$\bar{U} = \frac{1}{m}\sum_{j=1}^{m} U_j \quad (2)$$

e B é a variância entre imputação:

$$B = \frac{1}{m-1}\sum_{j=1}^{m}\left(Q_j - \bar{Q}_j\right)^2 \quad (3)$$

Então a estimativa de variância associada a $\bar{Q}$ é a variância total:

$$T = \bar{U} + \left(1 + \frac{1}{m}\right)B \quad (4)$$

Em seguida, foram construídos intervalos de confiança (95%) para a média ($\bar{Q}$) por meio de uma aproximação t-Student:

$$IC = \left[\bar{Q} \pm 1{,}96\sqrt{T}\right] \quad (5)$$

A eficiência relativa (ER) da IM de uma estimativa pontual baseada em m imputações foi quantificada pela Equação 6:

$$ER = 1 + \frac{FMI}{m}, \quad (6)$$

onde $FMI = \frac{B}{B+\bar{U}}$ é a fração de informações faltantes para a quantidade estimada (FMI) sobre Q, que varia de 0 a 1 (SCHAFER, 1997). A FMI quantifica a precisão da estimativa se não houvesse dados faltantes.

Rubin (1987) introduziu a fração de informação faltantes (λ) (7) para indicar o quanto as estimativas foram influenciadas pela presença de dados faltantes,

$$\lambda = \frac{B+\frac{B}{m}}{T} \quad (7)$$

O erro padrão (SE) da estimativa do parâmetro é dado por:

$$SE = \sqrt{\left(1 + \frac{\lambda}{m}\right)} \quad (8)$$

onde λ é a fração de informação ausente e m é o número de conjunto de dados completo.

2.2.4 Análise da eficiência da imputação

A eficiência do procedimento de imputação foi avaliada por meio da comparação dos parâmetros estimados (estimativas de parâmetros, erros padrão e intervalos de confiança de 95%), os coeficientes de determinação dos modelos de imputação e análises gráficas

(densidade de probabilidade e box-plot).

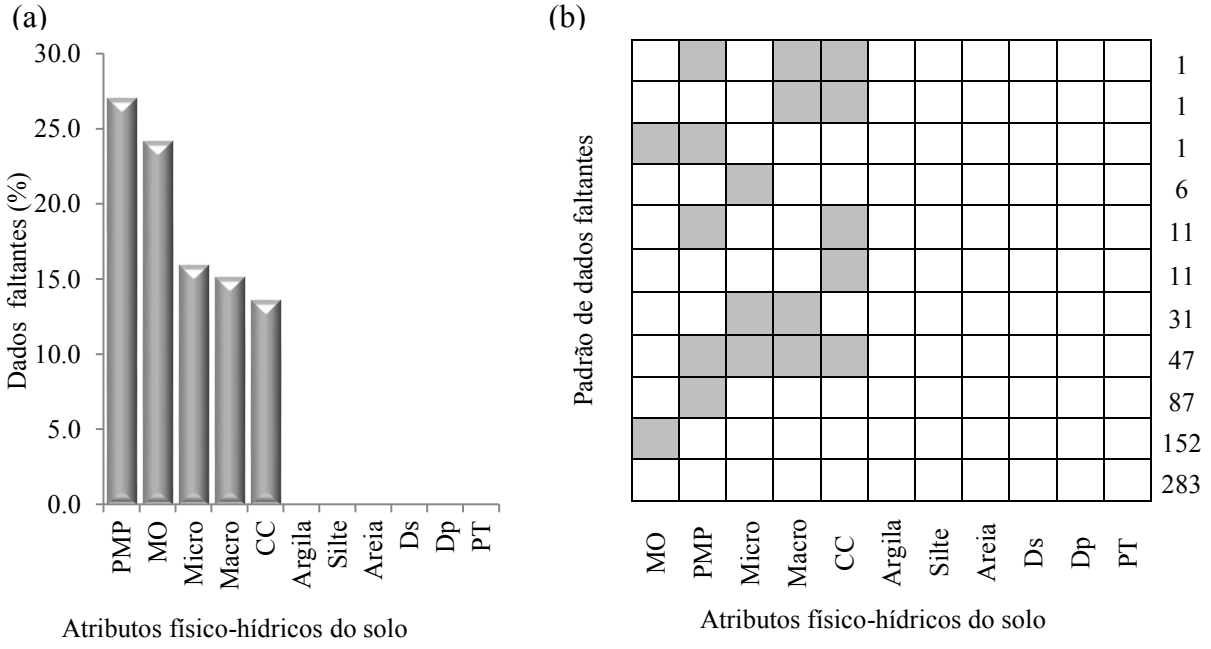
Todas as imputações múltiplas e análises foram realizadas no programa R (R CORE TEAM, 2017), utilizando o pacote MICE (Multivariate Imputation by Chained Equations).

2.3 Resultados

A análise preliminar para identificar a proporção e o padrão de dados perdidos é mostrada na Figura 1. A figura da esquerda mostra, em ordem decrescente, o percentual de valores faltantes para cada variável com dados omissos. Das onze variáveis que compõem o BDS, cinco têm dados faltantes, ou seja, 45%. A proporção de valores faltantes variou de 13,6 a 27,1%.

A Figura 1, à direita, exibe o padrão de dados faltantes onde as colunas são as variáveis e as linhas são as observações. Existem 283 amostras sem lacunas e 348 casos com dados faltantes, o que corresponde a um percentual de 55,2%. O padrão ausente foi multivariado e arbitrário. A matéria orgânica é a variável com maior quantidade de dados faltantes e tende a faltar em blocos de muitas observações.

Figura 1 – Gráfico de barras em ordem decrescente de valores faltantes (a) e padrão de dados faltantes (branco corresponde aos valores observados e valores faltantes em cinza) (b) de um banco de dados físico-hídrico do solo. (MO - matéria orgânica, PMP - ponto de murcha permanente, Micro - Microporosidade, Macro - macroporosidade, CC - Capacidade de campo, Ds - densidade do solo, Dp- densidade de partículas, PT - porosidade total).



Fonte: Próprio autor.

A partir do teste de Little (qui-quadrado, χ^2, de 850,89, com 91 graus de liberdade e p-valor = 0,000), pode-se afirmar, a um nível de significância de 5%, que os dados faltantes não são MCAR. Portanto, neste estudo, assumimos os dados do mecanismo MAR (ausente aleatoriamente).

Os resultados dos modelos preditivos (regressões lineares múltiplas) para as cinco propriedades do solo com dados perdidos (CC, PMP, Macro, Micro e MO) considerando a análise completa do caso (ACC) e a imputação múltipla (IM-MICE) estão resumidos na Tabela. 1. Os resultados completos do caso desviam-se notavelmente dos resultados imputados.

Onde os coeficientes são zero significa que essas variáveis têm pouca significância para prever a variável de interesse. A macroporosidade do solo foi o único parâmetro a ser atribuído quando nenhuma das variáveis resultou no coeficiente zero. Em geral, a densidade do solo foi de maior importância na predição das variáveis de desfechos.

A significância das covariáveis (valor de p) variou dependendo da variável a ser estimada (Tabela 1). Para a estimação do CC, as covariáveis mais relacionadas à estrutura do solo (Micro, Macro e Ds) foram significativas nos dois métodos. Por outro lado, para o modelo PWP, aqueles relacionados à textura (argila e areia) foram os mais relevantes. Para os modelos de estimação Macro, a significância variou de acordo com o método utilizado, apenas duas covariáveis (Micro e CC) foram mais expressivas para a estimação desta variável, quando o método ACC foi aplicado.

Tabela 1 - Parâmetros de estimativas por análise de caso completa (ACC) e imputação múltipla pelo método MICE (Imputação multivariada por equações encadeadas).

(Continua)

Covariáveis	Análise de Casos Completos (n=283)			Imputação – MCMC (n=631)				
	β[1] (Erro padrão)	IC (95%)[2]	P-valor[3]	β (Erro padrão)	IC [95%]	P-valor	Fmi[4]	λ[5]
				Capacidade de campo				
Intercepto	63,4(39,3)	[-13,7; 140,5]	0,108	59,4(14,3)	[31,1; 87,6]	0,00	0,1	0,1
Microporosidade	0,3(0,1)	[0,2; 0,4]	0,000**[6]	0,2(0,0)	[0,1; 0,3]	0,00**	0,2	0,2
Macroporosidade	-0,1(0,0)	[-0,2; 0,0]	0,004*[7]	-0,2(0,0)	[-0,2; -0,1]	0,00**	0,3	0,3
Porosidade total	-0,4(0,2)	[-0,7; 0,0]	0,051	-0,3(0,1)	[-0,5; -0,1]	0,017	0,2	0,2
Ponto de murcha permanente	0,1(0,1)	[-0,2; 0,3]	0,608	0,1(0,1)	[-0,1; 0,2]	0,385	0,3	0,3
Densidade do solo	-21,4(6,7)	[-34,5; -8,3]	0,002*	-14,9(4,0)	[-22,8; -7,0]	0,000**	0,2	0,2
Argila	0,2(0,3)	[-0,5; 0,8]	0,585	0,1(0,1)	[0,0; 0,3]	0,091	0,0	0,0
Silte	0,1(0,3)	[-0,6; 0,7]	0,849	0,1(0,1)	[0,0; 0,3]	0,068	0,0	0,0
Areia	00(0,3)	[-0,7; 0,6]	0,945	-0,1(0,1)	[-0,3; 0,0]	0,145	0,0	0,0
Matéria orgânica	00(0,4)	[-0,7; 0,8]	0,916	0,0(0,2)	[-0,3; 0,3]	0,963	0,3	0,2

[1] β = coeficientes
[2] IC (95%) = [intervalo de confiança inferior; intervalo de confiança superior]
[3] P-valor = probabilidade de se observar um valor da estatística de teste maior ou igual ao encontrado
[4] Fmi = fração de informação faltantes
[5] λ= proporção da variância total que é atribuível aos dados faltantes
[6] α (Nivel de significancia) =‘**’ 0,01
[7] α (Nivel de significancia) = ‘*’ 0,05

Tabela 1 - Parâmetros de estimativas por análise de caso completa (ACC) e imputação múltipla pelo método MICE (Imputação multivariada por equações encadeadas).

(Continuação)

Covariáveis	Análise de Casos Completos (n=283)			Imputação – MCMC (n=631)				
	β (Erro padrão)	IC (95%)	P-valor	β (Erro padrão)	IC [95%]	P-valor	Fmi	λ
				Ponto de murcha permanente				
Intercepto	28,7(7,0)	[15,0; 42,5]	0,000	32,2(7,8)	[16,1; 48,3]	0,000	0,4	0,4
Microporosidade	0,0(0,0)	[0,0; 0,1]	0,201	0,0(0,0)	[0,0; 0,1]	0,334	0,7	0,7
Macroporosidade	0,0(0,0)	[-0,1; 0,0]	0,265	0,0(0,0)	[-0,1; 0,0]	0,083	0,6	0,5
Capacidade de campo	0,0(0,0)	[0,0; 0,1]	0,135	0,0(0,0)	[0,0; 0,1]	0,398	0,2	0,2
Densidade da partícula	-2,2(1,7)	0,0;0,1]	0,196	-4,3(2,0)	[-8,5; -0,1]	0,045	0,6	0,5
Densidade do solo	2,1(0,9)	[-5,6; 1,1]	0,026*	2,2(0,9)	[0,3; 4,1]	0,024*	0,3	0,3
Argila	0,2(0,0)	[0,1; 0,3]	0,001**	0,2(0,1)	[0,1; 0,3]	0,000**	0,1	0,1
Silte	-0,1(0,0)	[-0,1; 0,0]	0,226	-0,1(0,0)	[-0,2; 0,0]	0,192	0,0	0,0
Areia	-0,3(0,0)	[-0,4; -0,2]	0,000**	-0,2(0,1)	[-0,3; -0,1]	0,000**	0,1	0,1
				Macroporossidade				
Intercepto	27,0(23,2)	[-18,6; 72,5]	0,247	17,2(20,0)	[-22,1; 56,5]	0,497	0,1	0,1
Microporosidade	-0,4(0,1)	[-0,5; -0,3]	0,000**	-0,4(0,1)	[-0,5; -0,3]	0,000**	0,5	0,5
Porosidade total	0,2(0,2)	[-0,1; 0,6]	0,241	0,4(0,2)	[0,1; 0,7]	0,005**	0,1	0,1
Ponto de murcha permanente	-0,1(0,1)	[-0,4; 0,1]	0,245	-0,3(0,1)	[-0,6; 0,0]	0,050*	0,6	0,6
Capacidade de campo	-0,2(0,1)	[-0,4; -0,1]	0,001**	-0,3(0,1)	[-0,5; -0,2]	0,000**	0,3	0,3
Densidade do solo	-12,5(6,7)	[-25,6; 0,6]	0,063	-4,5(5,3)	[-15,0; 6,0]	0,496	0,1	0,1
Argila	0,2(0,1)	[0,0; 0,4]	0,058	0,2(0,1)	[0,0; 0,5]	0,032*	0,0	0,0
Silte	0,1(0,1)	[-0,1; 0,3]	0,387	0,1(0,1)	[-0,1; 0,3]	0,343	0,0	0,0
Areia	0,2(0,1)	[-0,1; 0,4]	0,128	0,1(0,1)	[-0,1; 0,3]	0,359	0,0	0,0
Silte	0,1(0,3)	[-0,5; 0,7]	0,690	-0,2(0,1)	[-0,5; 0,0]	0,079	0,5	0,4
Areia	0,0(0,3)	[-0,6; 0,5]	0,901	-0,3(0,1)	[-0,6; 0,0]	0,035*	0,5	0,5
Matéria orgânica	1,5(0,3)	[0,8; 2,2]	0,000**	1,1(0,2)	[0,7; 1,5]	0,000**	0,4	0,3

Tabela 1 - Parâmetros de estimativas por análise de caso completa (ACC) e imputação múltipla pelo método MICE (Imputação multivariada por equações encadeadas).

(Conclusão)

Covariáveis	Análise de Casos Completos (n=283)			Imputação – MCMC (n=631)				
	β (Erro padrão)	IC (95%)	P-valor	β (Erro padrão)	IC [95%]	P-valor	Fmi	λ
				Microporosidade				
Intercepto	16,4(30,0)	[-9,1; 41,9]	0,585	44,0(13,2)	[16,2; 71,7]	0,004	0,5	0,5
Macroporosidade	-0,2(0,0)	[-0,2; -0,1]	0,000**	-0,3(0,0)	[-0,4; -0,2]	0,000**	0,4	0,4
Ponto de murcha permanente	0,2(0,1)	[0,0; 0,4]	0,074	0,0(0,1)	[-0,3; 0,4]	0,840	0,8	0,7
Capacidade de campo	0,3(0,1)	[0,2; 0,4]	0,000**	0,3(0,0)	[0,2; 0,4]	0,000**	0,2	0,2
Porosidade total	0,0(0,1)	[-0,1; 0,1]	0,454	0,1(0,1)	[0,0; 0,2]	0,118	0,6	0,5
Argila	0,0(0,3)	[-0,6; 0,6]	0,956	-0,2(0,1)	[-0,5; 0,0]	0,083	0,5	0,4
Silte	0,1(0,3)	[-0,5; 0,7]	0,690	-0,2(0,1)	[-0,5; 0,0]	0,079	0,5	0,4
Areia	0,0(0,3)	[-0,6; 0,5]	0,901	-0,3(0,1)	[-0,6; 0,0]	0,035*	0,5	0,5
Matéria orgânica	1,5(0,3)	[0,8; 2,2]	0,000**	1,1(0,2)	[0,7; 1,5]	0,000**	0,4	0,3
				Matéria Orgânica				
Intercepto	4,9(7,2)	[-9,2; 19,0]	0,495	-2,9(9,2)	[-22,6; 16,8]	0,756	0,6	0,5
Microporosidade	0,0(0,0)	[0,0; 0,1]	0,000**	0,1(0,0)	[0,0; 0,1]	0,014**	0,9	0,8
Porosidade total	-0,1(0,2)	[-0,4; 0,2]	0,521	0,2(0,2)	[-0,2; 0,6]	0,378	0,6	0,5
Capacidade de campo	0,0(0,0)	[0,0; 0,02]	0,751	0,00(0,0)	[-0,1; 0,1]	0,968	0,8	0,8
Densidade da partícula	2,2(3,3)	-4,3; 8,7]	0,501	-5,1(4,0)	[-13,8; 3,6]	0,225	0,6	0,5
Densidade do solo	-5,5(5,9)	[-17,1; 6,1]	0,356	4,8(7,3)	[-11,0; 20,7]	0,523	0,6	0,5
Silte	0,1(0,0)	[0,06; 0,1]	0,000**	0,0(0,0)	[0,0; 0,1]	0,000**	0,4	0,3
Areia	0,0(0,0)	[0,0; 0,03]	0,004**	0,0(0,0)	[0,0; 0,03]	0,047*	0,5	0,4

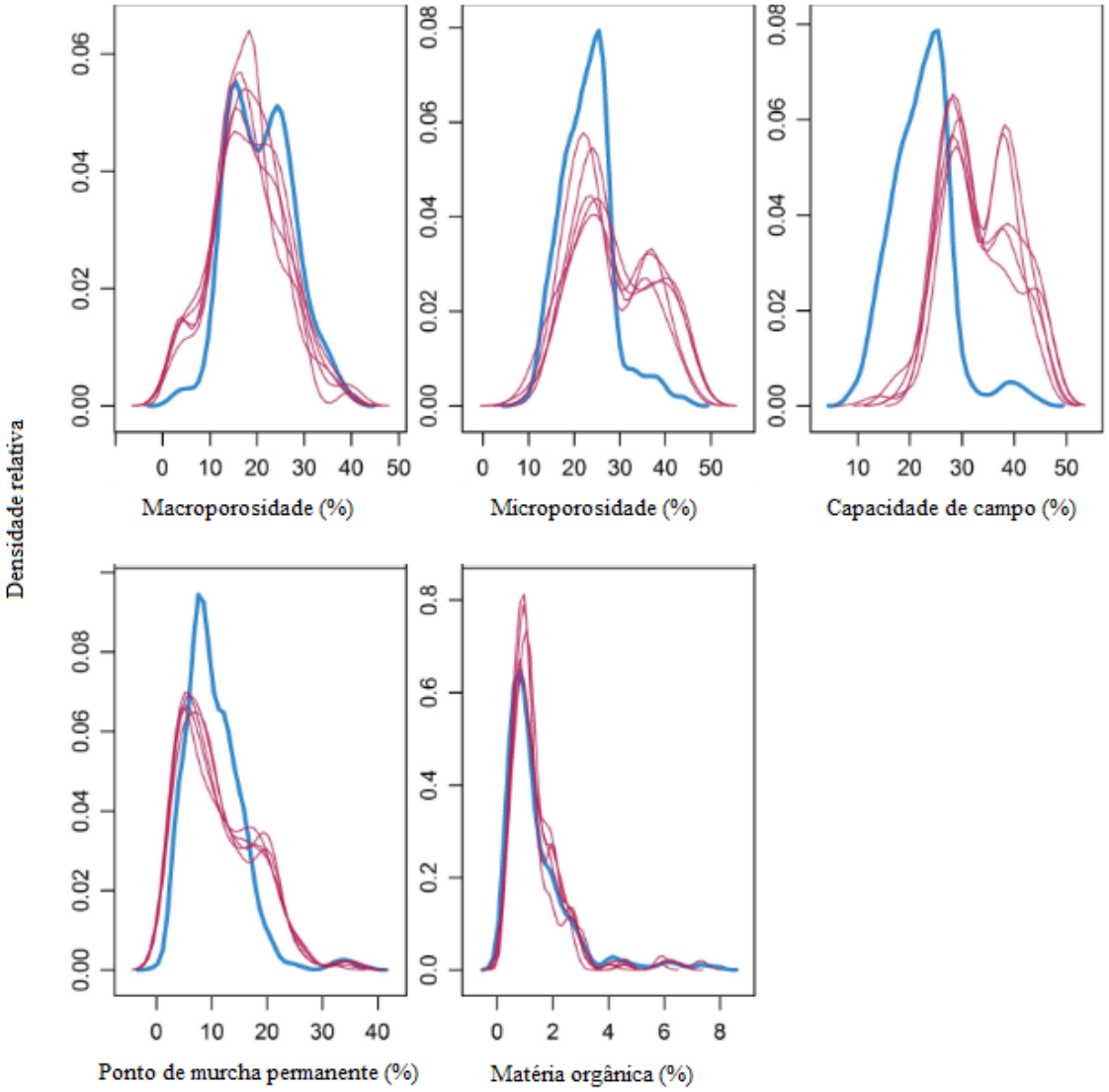
Fonte: Próprio autor.

A maior fração de informações faltantes (FMI) observada neste estudo foi na covariável Micro para estimar a MO (FMI = 0,9), ou seja, menor certeza estatística para estimar essa variável. A MO também foi superior na eficiência das estimativas (RP = 85%), com a fração de informação faltante (λ = 0,8) obtida com 5 imputações.

O erro padrão da estimativa do parâmetro foi $\sqrt{\left(1+\frac{0,8}{5}\right)} = 1,08$ vezes maior que o erro padrão com um número infinito de imputações. Vale destacar que as maiores proporções da variância total foram associadas às variáveis que apresentaram dados faltantes quando estas foram inseridas como covariáveis para estimar as demais.

A Figura 2 compara graficamente as distribuições das variáveis observadas (azul) e imputadas (vermelhas) nos modelos de imputação. As distribuições são muito semelhantes. A capacidade de campo (CC) e o ponto de murcha permanente (PMP) parecem desviar-se um pouco mais dos dados observados.

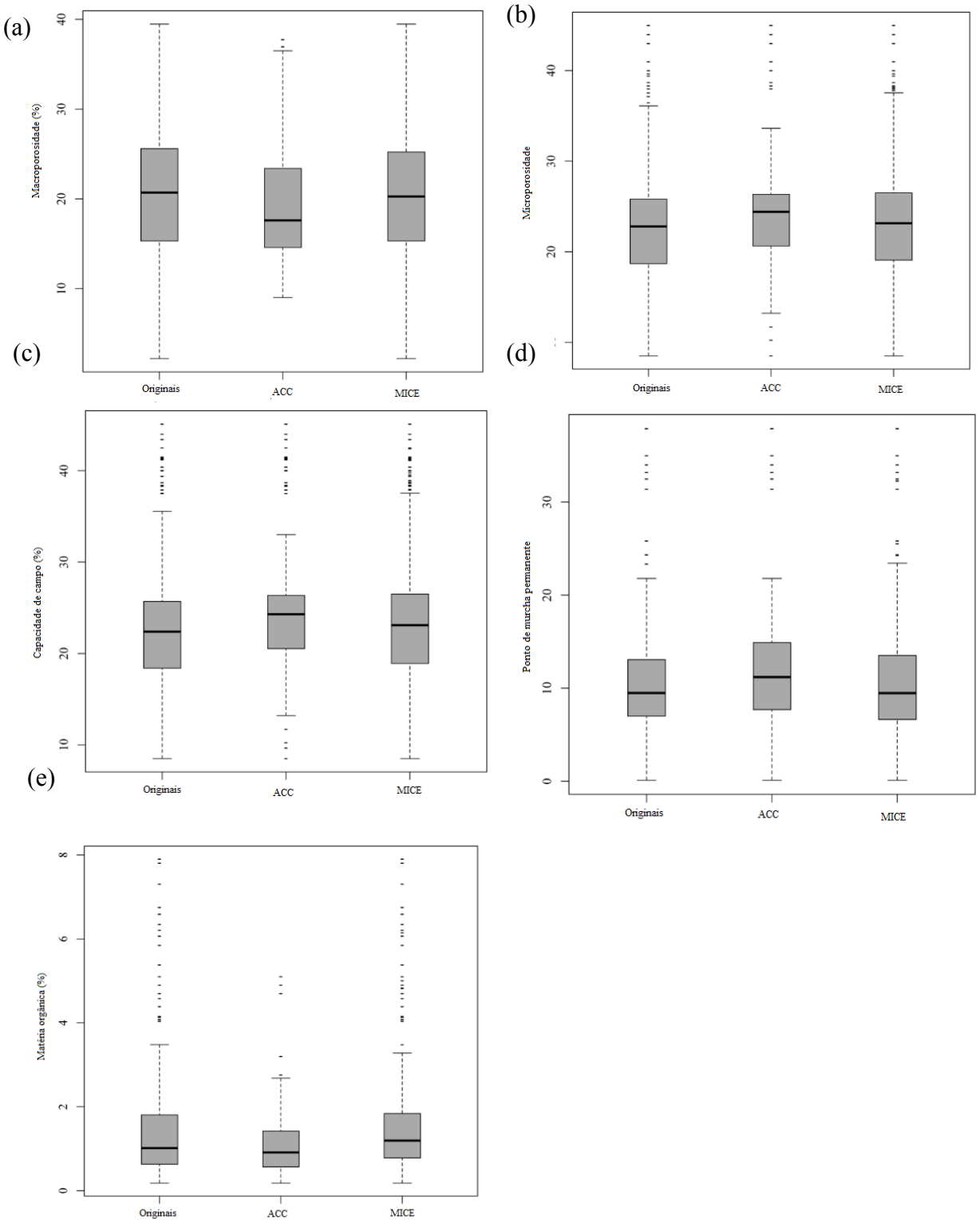
Figura 2 - Funções de densidade de probabilidade dos dados observados (linha azul) e as cinco cadeias geradas por MICE (linha vermelha) para as variáveis capacidade de campo (CC), ponto de murcha permanente (PMP), macroporosidade (Macro), microporosidade (Micro) e matéria orgânica (MO).



Fonte: Próprio autor.

A comparação da distribuição dos dados, para cada variável, tendo em conta os dados originais e submetidos à análise completa do caso (ACC) e após a imputação múltipla (MICE) é mostrada na Figura 3. Vale ressaltar como as múltiplas imputações via MICE mantiveram o mesmo comportamento que os dados originais, e mudança de distribuição foi observada para o CCA, especialmente para as variáveis Micro, CC e MO (Figuras 3b, 3c e 3e), onde as caixas centrais, que representam 50% dos dados, foram reduzido. As medianas foram maiores no ACC para Micro, CC e PMP.

Figura 3 - Box-plot das variáveis macroporosidade (a), microporosidade (b), capacidade de campo (c), ponto de murcha permanente (d) e matéria orgânica (e) com dados originais, múltipla imputação via MICE (Imputação Multivariada por Acorrentado Equações).



Fonte: Próprio autor.

2.4 Discussão

O exame inicial do SDB para o padrão de dados faltantes (Figura 1) é importante para a seleção do método de imputação a ser utilizado (HONAKER; KING; BLACKWELL, 2011). Segundo (FIGUEREDO et al., 2000) o problema da falta de dados na análise multivariada tem implicações que ameaçam a validade das conclusões.

Embora o ACC não seja um método de IM, é uma referência para verificar a variabilidade das estimativas (AUDIGIER et al., 2015). A maior proporção de variância atribuída a dados perdidos (FMI) foi observada para MO (Tabela 1) uma vez que essa covariante teve a maior taxa estimada de perda de informação (λ).

As covariáveis inseridas no modelo inicial (Tabela 1) foram todas aquelas que apresentaram correlação com a variável estimada. Buuren e Oudshoorn (2000) sugerem que o número de preditores utilizados para a imputação deve ser o mais amplo possível, uma vez que um grande conjunto de preditores tende a tornar a hipótese do MAR mais provável.

A convergência da amostragem MICE foi confirmada pelos gráficos de função de densidade de probabilidade (Figura 2), que apresentaram aproximadamente a mesma distribuição e, a similaridade das curvas confirma que o algoritmo de amostrador de Gibbs converge.

Por fim, a eficiência na imputação dos dados faltantes ficou evidente nos gráficos box-plots (Figura 3), pois o método de imputação MICE apresentou comportamento de distribuição dos dados semelhante aos dos dados observados, tanto assimetria quanto dispersão, em comparação aos dados submetidos a ACC. Ou seja, o método de IM preservou as características originais do BDS. As diferenças observadas nas ACC indicam que essa abordagem não permite generalizações para toda a população alvo de interesse.

Embora os resultados obtidos com a aplicação do MICE não tenham se destacado em relação a ACC (médias e desvios padrão semelhantes), a preservação da variabilidade original dos dados já demonstra que a aplicação da IM é uma alternativa apropriada para completar um banco de dados com valores faltantes, principalmente para análises multivariadas.

Na ciência do solo, essa situação pode ser exemplificada pela estimação das funções de pedotransferência, que obtidas a partir de uma abordagem multivariada, são usadas para estimar propriedades do solo que, ou são de determinação onerosa, ou não disponíveis (MINASNY; HARTEMINK, 2011) e, muitas vezes, os bancos de dados disponíveis têm lacunas, reduzindo consideravelmente o tamanho da amostra. Exemplificando para este fim os BDS utilizados neste estudo, o não preenchimento dos dados ausentes resultaria na redução

do BDS original de 631 para 283 e, consequentemente, na alteração da variabilidade dos dados, como demonstrado (Figura 1). Essa significativa modificação do banco de dados possivelmente resultaria em modelos diferentes daqueles obtidos para o banco completo, levando à imprecisão dos resultados.

Quando o número de casos disponíveis para análise multivariada é diminuído, o poder estatístico para detectar efeitos significativos é reduzido, levando potencialmente ao erro Tipo II (consiste na não-rejeição de uma hipótese inicial H_0, tida como falsa). As chances de erro do Tipo II aumentam quando a amostra original do estudo é pequena, como pode ocorrer em estudos experimentais que avaliam a eficácia do tratamento. O principal problema na eliminação de listas (ACC) é se o tamanho restante da amostra é suficiente para fornecer poder estatístico adequado, uma vez que dados ausentes podem causar a exclusão de grande parte dos dados originais (FIGUEREDO et al., 2000).

Apesar dos avanços metodológicos e demonstrações da eficiência do IM em diversas áreas (SQUILLANTE et al., 2018) na Ciência do Solo, essa abordagem ainda é subutilizada para lidar com dados perdidos. Este trabalho evidenciou as vantagens desta técnica na estimativa de dados de propriedades físico-hídricas do solo. Portanto, os resultados observados aqui podem ser usados em estudos com conjuntos de dados semelhantes. Nesse caso, recomendamos que o método IM-MICE seja preferido em relação ao ACC. Já que analisar apenas os casos completos, resulta em amostras menores, ou seja, perda de informação, com menor precisão estatística nas análises (NUNES; KLÜCK; FACHEL, 2009).

2.5 Conclusões

1. A imputação múltipla por equação de cadeia (MICE) apresentou desempenho satisfatório, pois preservou as características dos dados originais como, a distribuição, simetria e a dispersão dos dados.
2. A IM é um método muito importante para lidar com dados faltantes. Com este estudo, pretende-se ajudar mais pesquisadores do solo a começar a implementar técnicas de IM, como MICE, em vez de excluir amostras com dados faltantes e, com isso mantendo o tamanho de seu banco de dados.

REFERÊNCIAS

AUDIGIER, V.; HUSSON, F.; JOSSE, J. Multiple imputation for continuous variables using a Bayesian principal component analysis. **Journal of Statistical Computation and Simulation,** v. 86, p. 2140-2156, 2015.

BUUREN, S. **Flexible Imputation of Missing Data**. 2. ed. Chapman and Hall/CRC, 2018, 416 p.

CARVALHO, J.R.P MONTEIRO, J.E.B.A.; NAKAI, A.M.; ASSAD, E.D. Modelo de imputação múltipla para estimar dados de precipitação diária e preenchimento de falhas. **Revista Brasileira de Meteorologia**, v. 32, p. 575–583, 2017.

CLAESSEN, M.E.C.; BARRETO, W.O.; PAULA, J.L.; DUARTE, M.N. **Manual de métodos de análise de solo**. Rio de Janeiro: EMBRAPA, 1997. 212 p.

CLIFFORD, D.; DOBBIE, M.J.; SEARLE, R. Non-parametric imputation of properties for soil profiles with sparse observations. **Geoderma**, v. 232–234, p. 10–18, 2014.

FIGUEREDO, A.J.; MCKNIGHT, P.E.; MCKNIGHT, K.M.; SIDANI, S. Multivariate modeling of missing data within and across assessment waves. **Addiction,** v. 95, p. 361–380, 2000.

GRAHAM, J.W. Missing data analysis: Making it work in the real world. **Annual Review of Psychology**, v. 60, p.549–576, 2009.

HARRELL, J.F.E. **Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis**. 2. ed. New York: Springer International Publishing; 2016. 572 p.

HONAKER, J.; KING, G.; BLACKWELL, M. Amelia II: A program for missing data. **Journal of Statistical Software,** v. 45, p. 1–47, 2011.

KIM, M.; BAEK, S.; LIGARAY, M.; PYO, J. ;PARK, M.; CHO, K.H. Comparative studies of different imputation methods for recovering streamflow observation. **Water Resource Research,** v. 7, p. 6847–6860, 2015.

LITTLE, R.J.; RUBIN, D.B. Missing Data. **International Encyclopedia of the Social and Behavioral Sciences**, v. 15, p. 602-607, 2015.

LITTLE, R.J.A. A Test of Missing Completely at Random for Multivariate Data with Missing Values. **Journal of the American Statistical Association**, v. 83, p. 1198–1202, 1988.

NUNES, L.N.; KLÜCK, M.M.; FACHEL, J.M.G. Uso da imputação múltipla de dados faltantes: uma simulação utilizando dados epidemiológicos. **Caderno de Saúde Pública**, v. 25, n.2, p. 268-278, 2009.

MINASNY, B.; HARTEMINK, A.E. Predicting soil properties in the tropics. **Earth-Science Reviews**, v. 106, p. 52–62, 2011.

PAES, Â.T.; POLETO, F.Z. Por dentro da estatística. Educ Contin Saúde Einstein. 2013; 11:5-7.

PEDERSEN, A.B.; MIKKELSEN, E.M.; CRONIN-FENTON, D.; KRISTENSEN, N. R.; PHAM, T.M. PEDERSEN, L.; PETERSEN, I. Missing data and multiple imputation in clinical epidemiological research. **Clinical Epidemiology**, v. 9, p. 157-166, 2017.

POYATOS, R.; SUS, O.; BADIELLA, L.; MENCUCCINI1, M.; MARTÍNEZ-VILALT, J. Gap-filling a spatially explicit plant trait database: comparing imputation methods and different levels of environmental information. **Biogeosciences**, v. 15, p. 2601–2617, 2018.

R CORE TEAM (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. [accessed on 15 January 2018].

RUBIN, D.B. Inference and missing data. **Biometrika**, v. 63, p. 581–592, 1976.

RUBIN, D.B. **Multiple Imputation for Nonresponse in Surveys**. 1. ed. New York: John Wiley & Sons, 1987. 253 p.

SANTOS, H.G. *et al*. **Sistema brasileiro de classificação de solos**. 3. ed. rev. ampl. Rio de Janeiro: Embrapa Solos; 2013. 353 p.

SCHAFER, J.L.; OLSEN, M.K. Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. **Multivariate Behavioral Research**, v. 33, p. 545-571, 1998.

SHAO, J.; MENG, W.; SUN, G. Evaluation of missing value imputation methods for wireless soil datasets. **Personal and Ubiquitous Computing**, v. 21, p. 113–123, 2017.

SILVA, A.C.; ARMINDO, R.A. Importância das funções de pedotransferência no estudo das propriedades e funções hidráulicas dos solos do Brasil. **Multi-Science Journal,** v. 1, p. 31–37, 2016.

SONG, Q.; SHEPPERD, M. A new imputation method for small software project data sets. **Journal of Systems and Softwares**, v. 80, p. 51–62, 2007.

SQUILLANTE, Jr.R.; FO, D.J.S.; MARUYAMA, N.; JUNQUEIRA, F.; MOSCATO, L. A.; NAKAMOTO, F. Y.; MIYAGI, P.E.; OKAMOTO, Jr.J. Modeling accident scenarios from databases with missing data: A probabilistic approach for safety-related systems design. **Safety Science**, v. 104, p. 119–134, 2018.

STUART, E.A.; AZUR, M.; FRANGAKIS, C.; LEAF, P. Multiple imputation with large data sets: a case study of the Children's Mental Health Initiative. **American Journal of Epidemiology**, v. 169, n. 9, p.1133–1139, 2009.

VAN BUUREN, S.; OUDSHOORN, C.G.M. **Multivariate Imputation by Chained Equations: MICE V1.0 User´s manual**. Leiden: TNO Preventie en Gezondheid, TNO/PG/VGZ/00.038; 2000.

3 FUNÇÕES DE PEDOTRANSFERÊNCIA PARA ESTIMAR A DENSIDADE DO SOLO NA AMAZÔNIA ORIENTAL, BRASIL

RESUMO

Funções de pedotransferência (FPTs) são equações utilizadas para modelar atributos do solo de difícil obtenção a partir de dados disponíveis. As FPTs tem sido muito utilizadas para predizer a densidade do solo (Ds), atributo importante na avaliação da qualidade estrutural do solo, entretanto, não é geralmente encontrada em banco de dados de solo, principalmente quando as informações são provenientes de grandes regiões, como é o caso da região amazônica. Objetivou-se com este estudo: (i) desenvolver FPTs para estimar a densidade do solo na Amazônia oriental; e (ii) comparar o desempenho das FPTs com modelos disponíveis na literatura. Um procedimento "stepwise" foi utilizado na seleção de variáveis preditoras. Três novas FPTs foram desenvolvidas para predição da Ds para diferentes regiões do estado do Pará, Brasil. A técnica de regressão linear múltipla foi aplicada para validar as equações usando um conjunto de dados independente. Três FPTs provenientes da literatura foram comparadas e avaliadas. O melhor desempenho foi obtido para a FPT1, que tem como variáveis preditoras o teor de silte, a densidade de partícula e o carbono orgânico do solo. As equações desenvolvidas mostraram melhor precisão na previsão da Ds em comparação com as equações já existentes.

Palavras-chave: Regressão linear múltipla. Stepwise. Modelo de validação

ABSTRACT

Pedotransfer functions (FPTs) are equations used to model soil attributes that are difficult to obtain from available data. FPTs have been widely used to predict soil density (Ds), an important attribute in the assessment of soil structural quality, however, it is not generally found in soil databases, especially when information comes from large regions such as This is the case of the Amazon region. The objective of this study was: (i) to develop FPTs to estimate soil density in eastern Amazonia; and (ii) compare the performance of PTFs with models available in the literature. A stepwise procedure was used to select predictor variables. Three new FPTs were developed to predict DS for different regions of the state of Pará, Brazil. The multiple linear regression technique was applied to validate the equations using an independent data set. Three FPTs from the literature were compared and evaluated. The best performance was obtained for FPT1, which has as predictor variables silt content, particle density and soil organic carbon. The developed equations showed better accuracy in predicting Ds compared to existing equations.

Keywords: Multiple linear regression. Stepwise Model validation.

3.1 Introdução

A densidade do solo (Ds) é um atributo físico importante no prognóstico da retenção e movimento da água no solo, estoque de carbono e avaliação de camadas compactadas (SUUSTER et al., 2011). No entanto, apesar da importância da Ds, a mesma não é frequentemente mensurada nos levantamentos de solos (SUUSTER et al., 2011) por demandar tempo e mão de obra (CHEN et al., 2018). Consequentemente, não estão rotineiramente disponíveis em banco de dados amplos, principalmente quando provenientes de regiões extensas, como é o caso da Região Amazônica.

Uma alternativa para lidar com a falta de informações de Ds em é o uso de funções de pedotransferência (FPTs), as quais têm sido utilizadas como uma opção à medição direta. Com uso destas funções, a Ds pode ser estimada a partir de atributos de solo disponíveis ou, mais facilmente medidos, geralmente encontrados nos bancos de dados de solos (CHEN et al., 2018). O uso de FPTs é baseado na premissa de que a Ds é fortemente correlacionada com outros atributos elementares do solo, tais como: textura, carbono orgânico, densidade de partículas, e profundidade do solo (TOMASELLA e HODNETT, 1998; BERNOUX et al., 1998; SEQUEIRA et al., 2014).

Nas últimas décadas, FPTs para estimar a Ds foram desenvolvidas para diversas localidades em diferentes escalas e métodos (BEUTLER et al., 2017; NGUYEN et al., 2017; QIAO et al., 2018). Todavia, FPTs desenvolvidas para uma determinada região, quando aplicadas a ambientes diferentes, podem fornecer estimativas imprecisas (NANKO et al., 2014). Desta forma, é mais seguro utilizar uma FPT desenvolvida a partir de dados do local de aplicação ou para uma área com solos de gênese semelhante (NEMES et al., 2010).

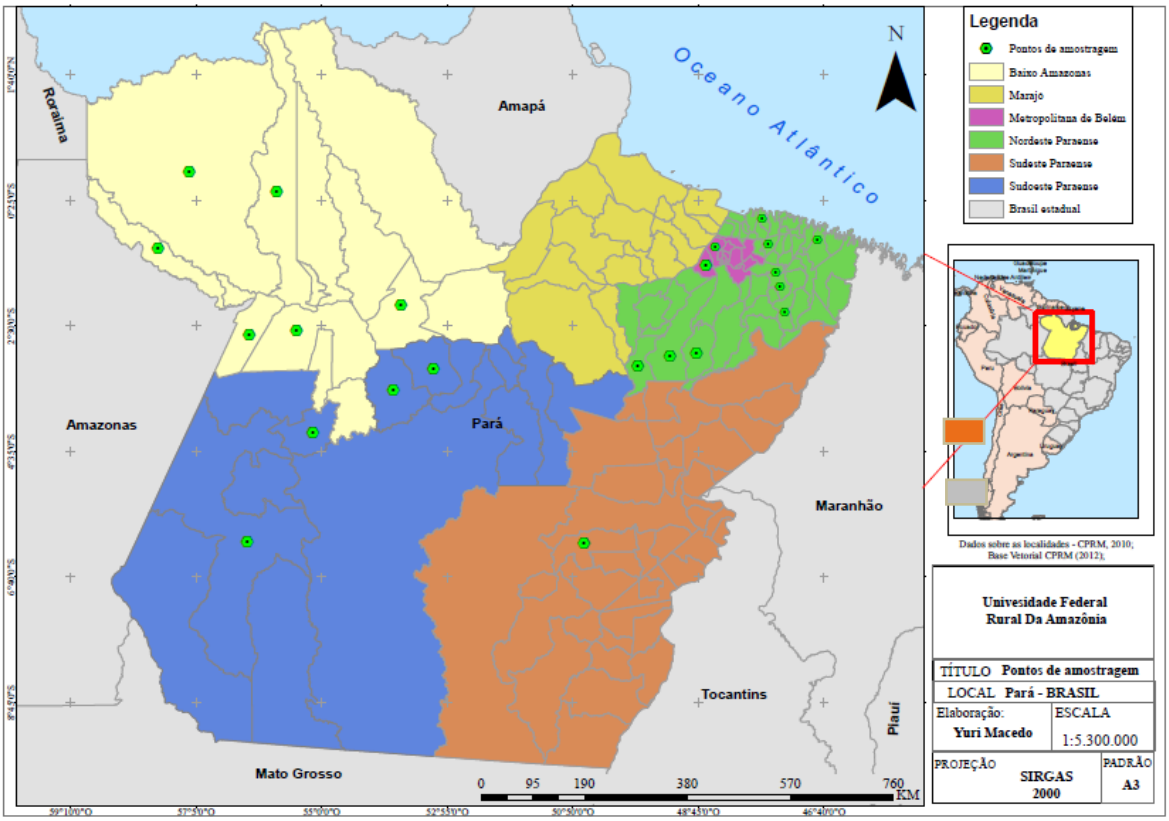
Neste contexto, os objetivos deste estudo foram: i) desenvolver PTFs para estimar a densidade do solo na Amazônia oriental; e ii) comparar o desempenho das PTFs com modelos disponíveis na literatura.

3.2 Material e métodos

3.2.1 Banco de dados

Foram utilizadas informações de solos de 23 municípios do Estado do Pará, região norte do Brasil (Figura 1). O estado é um dos nove estados que compõe a Amazônia Legal. O clima da região é equatorial, com temperatura média anual entre 24° e 26° C, com altos índices pluviométricos, que chegam a alcançar 2.000 mm (ALVARES et al., 2013).

Figura 1 - Localização das áreas de estudo (Municípios estudados por região do Estado do Pará).



Fonte: Próprio autor.

Os dados de solos foram obtidos a partir de Boletins de Pesquisa e Desenvolvimento da Embrapa, levantamentos de solos realizados pela Embrapa Amazônia Oriental, dissertações, teses e artigos científicos, realizados entre os anos de 1997 a 2014. Foram selecionadas amostras que continham dados da densidade do solo de duas classes de solos: Latossolos e Argissolos, classificados segundo Santos et al. (2018). O banco de dados contém informações de localização da amostragem, classificação do solo, profundidade da coleta, composição granulométrica (teores de areia, silte e argila), densidade do solo (Ds), densidade de partículas (Dp), macro porosidade (Macro), micro porosidade (Micro), porosidade total (PT), capacidade de campo (CC), ponto de murcha permanente (PMP) e conteúdo de carbono orgânico do solo (CO). Para a composição do banco de dados utilizado neste trabalho foram selecionadas somente as informações provenientes das camadas de 0 a 60 cm de profundidade.

3.2.2 Análise exploratória dos dados

Inicialmente foram verificados os limiares dos dados (mínimo e máximo admissível) para cada propriedade do solo e padronização de unidades de medidas (% e g cm^{-3}). A análise exploratória dos dados foi realizada por meio da estatística descritiva incluindo: máximo, mínimo, mediana, média, desvio padrão e o coeficiente de variação (CV).

3.2.3 Desenvolvimento das funções de pedotransferência (FPTs)

As funções de pedotransferência foram desenvolvidas de acordo com as seguintes etapas:

1º) Divisão aleatória dos dados em dois subconjunto: a) 85% dos dados para desenvolvimento e b) 15% para validação;

2º) Desenvolvimento das funções de pedotransferência considerando todas as classes de solo (Geral = Argissolos + Latossolos);

3º) Desenvolvimento de funções de pedotransferência por classes de solo (Argissolos e Latossolos.

O número de dados utilizados está apresentado na Tabela 1.

Tabela 1 – Número de dados utilizados para o desenvolvimento e validação das funções de pedotransferência.

Classes de solos	Total	Desenvolvimento	Validação
Geral	471	401	70
Argissolos	75	64	11
Latossolos	396	337	59

Fonte: Próprio autor.

As variáveis independentes foram consideradas no modelo por regressão linear múltipla pelo método stepwise com opção “direction=both", usando o Critério de Informação de Akaike (AIC) (AKAIKE, 1973; CARRERA e NEUMAN, 1986; MOLDRUP et al., 2004):

$$AIC = n\left[\ln(2\pi) + \ln\left(\frac{\sum_{i=1}^{n} d_i^2}{n-K}\right) + 1\right] + K, \text{ em que} \qquad (2)$$

ln representa o logaritmo natural e K é o número de parâmetros do modelo. Segundo Minasny et al. (1999), um valor AIC menor (ou mais negativo) indica melhor desempenho do modelo.

3.2.4 Avaliação das funções de pedotransferência (FPTs)

A avaliação do desempenho dos modelos de regressão foi realizada graficamente entre a relação dos valores estimados verso valores medidos e pelas análises dos resíduos. E utilizado indicadores estatísticos como o coeficiente de determinação (R^2), coeficiente de determinação ajustado (R_{aj}^2), erro médio a (EM), raiz do quadrado do erro médio (RMSE), obtidos pelas Equações 3, 4, 5 e 6 respectivamente. Foi realizado ainda o índice de desempenho (C) pela Equação 9, obtido pelo produto do índice de concordância (d) de Willmott (1981), (Equação 7) e do coeficiente de correlação de Pearson (r), Equação 8. O C foi classificado de acordo com Camargo e Sentelhas (1997) (Tabela 1).

$$R^2 = \frac{\left[\sum_{i=1}^{n}(\mathrm{Ds_{pi}}-\overline{\mathrm{Ds}}_{\mathrm{pi}})(\mathrm{Ds_{mi}}-\overline{\mathrm{Ds}}_{\mathrm{mi}})\right]^2}{\sum_{i=1}^{n}(\mathrm{Ds_{pi}}-\overline{\mathrm{Ds}}_{\mathrm{pi}})^2 \sum_{i=1}^{n}(\mathrm{Ds_{mi}}-\overline{\mathrm{Ds}}_{\mathrm{mi}})^2} \quad (3)$$

$$R_{aj}^2 = 1 - \left[(1-R^2) * \frac{n-1}{n-p}\right] \quad (4)$$

$$\mathrm{EM} = \frac{\sum_{\mathrm{i=1}}^{\mathrm{n}}(\mathrm{Ds_{mi}}-\mathrm{Ds_{pi}})^2}{\mathrm{n}} \quad (5)$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{\mathrm{i=1}}^{\mathrm{n}}(Ds_{\mathrm{mi}}-\mathrm{Ds_{pi}})^2}{\mathrm{n}}} \quad (6)$$

$$\mathrm{d} = 1 - \frac{\sum_{\mathrm{i=1}}^{\mathrm{n}}(\mathrm{Ds_{pi}}-\mathrm{Ds_{mi}})^2}{\sum_{\mathrm{i=1}}^{n}(|\mathrm{Ds_{pi}}-\overline{\mathrm{Ds}}_{\mathrm{mi}}|+|\mathrm{Ds_{mi}}-\overline{\mathrm{Ds}}_{\mathrm{mi}}|)^2} \quad (7)$$

$$r = \frac{\sum_{i=1}^{n}(\mathrm{Ds_{pi}}-\overline{\mathrm{Dds}}_{\mathrm{pi}})(\mathrm{Ds_{mi}}-\overline{\mathrm{Ds}}_{\mathrm{mi}})}{\sqrt{\left[\sum_{i=1}^{n}(\mathrm{Ds_{pi}}-\overline{\mathrm{Ds}}_{\mathrm{pi}})^2\right]\left[\sum_{i=1}^{n}(\mathrm{Ds_{mi}}-\overline{\mathrm{Ds}}_{\mathrm{mi}})^2\right]}} \quad (8)$$

$$C = r * d \quad (9)$$

onde: d pode assumir valores entre 0 e 1, sendo que d=1 indica uma perfeita concordância e d=0 uma total discordância entre os valores observados e preditos. Ds_{mi} é a densidade do solo medida (g cm $^{-3}$), Ds_{pi} é a densidade do solo predita (g cm^{-3}), $\overline{Ds_m}$ representa a média da densidade do solo medida e n indica o número total de observações, p é o número de parâmetros do modelo, incluindo o intercepto.

Tabela 2 - Critério de interpretação do desempenho dos modelos de regressão pelo índice de desempenho (c) proposto por Camargo e Sentelhas (1997).

Valor de C	Desempenho
≥ 0,85	Ótimo
0,75 --|0,85	Muito Bom
0,65 --| 0,75	Bom
0,60 --| 0.65	Mediano
0,50 --| 0,60	Sofrível
0,40 --| 0,50	Mau
< 0,40	Péssimo

Fonte: Camargo e Sentelhas (1997).

Os dados foram submetidos aos testes dos pressupostos necessários à validação do modelo de regressão linear: normalidade dos resíduos, variância constante dos resíduos (homocedasticidade) e independência entre os resíduos. A verificação dos pressupostos foi realizada por meio de análises gráficas e de um conjunto de testes simultâneos ou globais (PEÑA e SLATE, 2006), utilizando o pacote "gvlma" no software R (PEÑA e SLATE, 2019).

3.2.5 Comparação das Funções de pedotransferência (FPTs) desenvolvidas com as existentes na literatura

O desempenho de três FPTs para estimar a Ds, previamente publicadas na literatura (Tabela 1), foi comparado aos modelos desenvolvidos neste estudo com base nos valores de R^2_{aj} (Equação 4), EM (Equação 5) e RMSE (Equação 6) utilizando o banco de dados de validação. Os estudos de Bernoux et al. (1998) e Tomasella e Hodnett (1998), foram selecionados por apresentar modelos para estimativa da Ds a partir de dados de solos da Amazônia brasileira, enquanto, Benites et al. (2007) apresenta um modelo para estimar a Ds em diferentes regiões do Brasil.

Tabela 3 - Funções de pedotransferência selecionadas para a previsão da Ds.

Autores	Região	Modelo	n[8]	R^2_{aj}[9]
Benites et al.(2007)	Brasil	Ds = 1,5688 – 0,0005Argila-0,009COS[10]	1396	0,63
Bernoux et al. (1998)	Amazônia brasileira	Ds = 1,398 – 0,0047 Argila – 0,042COS	323	0,50
Tomasella e Hodnett (1998)	Amazônia brasileira	Ds = 1,578 – 0,054 COS – 0,006 Silte − 0,004 Argila	396	0,60

Fonte: Adaptado de Benites et al. (2007).

Todas as análises estatísticas foram realizadas usando o software R versão 3.5.1 (R CORE TEAM, 2018).

3.3 Resultados e Discussão

3.3.1 Análise descritiva dos dados

A estatística descritiva dos atributos físicos do solo para todos os solos (Geral) e por classe de solo para os dados de desenvolvimento das funções de pedotransferência (FPTs) estão apresentadas na Tabela 3. O valor médio da Ds dos dados gerais foi de 1,47 g cm^{-3}, variando entre 0,86 e 1,74 g cm^{-3}, no subconjunto de dados de Argissolos a média da Ds foi reduzida para 1,38 g cm^{-3}, variando entre 0,96 e 1,71 g cm^{-3} e nos Latossolos foi de 1,49 g cm^{-3}, variando entre 0,86 e 1,74 g cm^{-3}. O valor mediano foi um pouco mais elevado nos Latossolos.

Nos dados gerais e nos argissolos, todos os atributos do solo, exceto as medições de Dp, tiveram coeficiente de variação (CV > 10%). E nos Latossolos além da Dp, a Ds também apresentou baixo coeficiente de variação (CV < 10%). A Ds teve correlação positiva e significativas com a Dp e com o conteúdo de areia (Tabela 3). No entanto, a Ds foi negativamente correlacionada com o CO e teor de silte. Para os Argissolos a Ds também foi negativamente correlacionada com o teor de argila. A correlação mais forte da Ds foi com o silte e como o COS, o que indica a maior contribuição do teor de silte e do COS para a Ds em comparação com os outros atributos do solo nesta área de estudo.

[8] n = número de amostras
[9] R^2_{aj}=coeficiente de determinação ajustado
[10] COS = Carbono orgânico do solo

Tabela 4 - Estatística descritiva da variável argila, silte, areia, carbono orgânico do solo (CO), densidade do solo (Ds) e densidade de partículas (Dp) para todos os solos (Geral) e por classe de solos para os dados de desenvolvimento.

Estatísticas Descritivas	Argila	Silte	Areia	COS	Ds	Dp
	----------(%)----------				---------(g cm-^{3})-------	
			Geral (n = 401)			
Mínimo	3.00	1.70	11.00	0.10	0.86	2.22
Máximo	75.00	64.00	93.00	4,11	1.74	2.89
Mediana	16.70	10.00	70.00	0.53	1.49	2.63
Média	18.00	13.11	68.89	0.71	1.47	2.62
Desvio padrão	10.37	9.96	14.41	0.58	0.15	0.08
CV(%)[11]	57.57	76.00	20.92	81.69	10.11	3.02
r[12]	-0.09 ^{ns}	-0.46*	0.39*	-0.42*	1.00	0.28*
			Argissolos (n = 64)			
Mínimo	4.00	3.00	11.00	0.26	0.96	2.40
Máximo	72.00	51.00	91.10	4,11	1.71	2.72
Mediana	14.00	14.50	71.90	0.74	1.40	2.60
Média	16.67	16.94	66.40	0,91	1.38	2.59
Desvio padrão	11.60	10.91	16.84	0.70	0.16	0.06
CV(%)	69.58	64.40	25.36	76.92	11.60	2.38
r	-0.19*	-0.42*	0.40*	-0.43*	1.00	0.46*
			Latossolos (n = 337)			
Mínimo	3.00	1.70	13.00	0.10	0.86	2.22
Máximo	75.00	64.00	93.00	4.11	1.74	2.89
Mediana	17.00	10.00	70.00	0.53	1.51	2.63
Média	18.26	12.38	69.36	0.68	1.49	2.62
Desvio padrão	10.11	9.62	13.88	0.55	0.14	0.08
CV(%)	55.39	77.69	20.01	80.72	9.38	3.09
r	-0.09^{ns}	-0.45*	0.37*	-0.38*	1.00	0.22*

Fonte: Próprio autor.

Os dados de desenvolvimento (Tabela 3) e validação (Tabela 4) apresentaram poucas características contrastantes. Os dados de desenvolvimento tiveram conteúdo de COS, com valores máximos atingindo 4,11%, enquanto que nos dados de validação o valor máximo do conteúdo de OCS foi de 1,68%. Nos dados de desenvolvimento a Ds do solo foi negativamente correlacionada apenas com o teor de argila nos Argissolos, enquanto que nos dados de validação foi negativamente correlacionada em todos os subconjuntos de dados, apresentando maior correlação nos Argissolos.

[11] CV = coeficiente de variação
[12] r = coeficiente de correlação linear de Pearson

Tabela 5 - Estatística descritiva da variável argila, silte, areia, carbono orgânico do solo (CO), densidade do solo (Ds) e densidade de partículas (Dp) para todos os solos (Geral) e por classe de solo para os dados de validação.

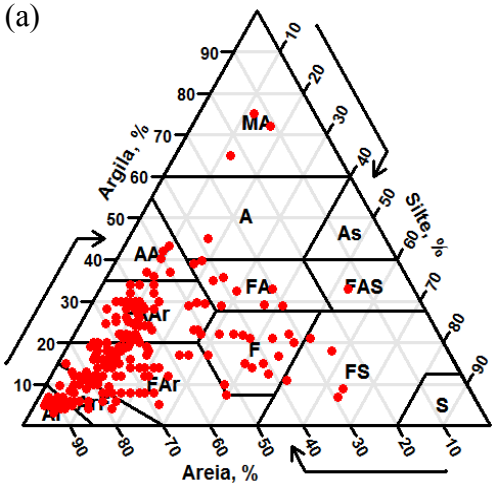
Estatísticas Descritivas	Argila	Silte	Areia	COS	Ds	Dp
	----------(%)----------				---------(g cm-^{3})--------	
			Geral (n = 70)			
Mínimo	4.00	4.00	13.00	0.12	0.84	2.43
Máximo	70.00	54.00	89.00	3.39	1.68	2.73
Mediana	18.00	10.00	68.00	0.58	1.49	2.62
Média	20.80	13.76	65.44	0.74	1.45	2.62
Desvio padrão	14.02	9.82	17.64	0.62	0.18	0.06
CV(%)	67.42	71.37	26.96	82.74	12.76	2.37
r	-0.28*	-0.63*	0.57*	-0.54*	1.00	0.27*
			Argissolos (n = 11)			
Mínimo	6.00	4.80	13.00	0.39	0.93	2.53
Máximo	40.00	47.00	86.00	1.86	1.51	2.67
Mediana	13.30	14.00	74.00	0.64	1.39	2.61
Média	17.75	18.80	63.45	0.98	1.34	2.60
Desvio padrão	12.40	13.07	24.53	0.55	0.17	0.05
CV(%)	69.84	69.54	38.66	55.90	12.74	1.95
r	-0.77*	-0.84*	0.84*	-0.72*	1.00	0.00^{ns}
			Latossolos (n = 59)			
Mínimo	4.00	4.00	19.00	0.12	0.84	2.43
Máximo	70.00	54.00	89.00	3.39	1.68	2.73
Mediana	18.70	10.00	68.00	0.53	1.52	2.63
Média	21.37	12.82	65.81	0.70	1.47	2.62
Desvio padrão	14.33	8.92	16.30	0.62	0.18	0.06
CV(%)	67.07	69.57	24.77	88.82	12.33	2.43
r	-0.25*	-0.57*	0.53*	-0.50*	1.00	0.29*

Fonte: Próprio autor.

As figuras 2a e b apresentam a distribuição textural dos conjuntos de dados de desenvolvimento e validação, respectivamente, para todos os solos (Geral) e por classes de solos. Segundo a classificação de Lemos e Santos (1996), as classes texturais predominantemente foram franco-argilo-arenosa, franco-arenosa, argilo-arenosa e areia franca.

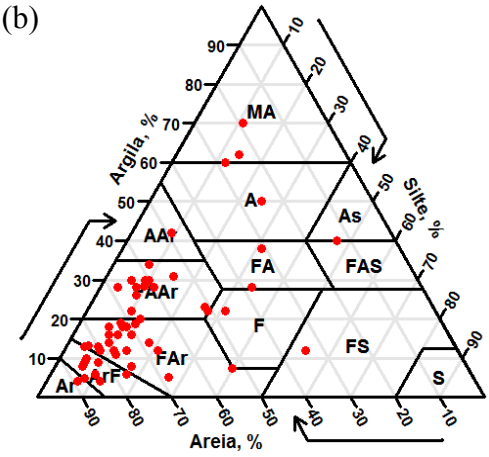
Figura 2 - Distribuição textural de todos os solos (Geral) e por classes de solos da Amazônia Oriental utilizadas no desenvolvimento (a) e validação (b) da função de pedotransferência.

(a)

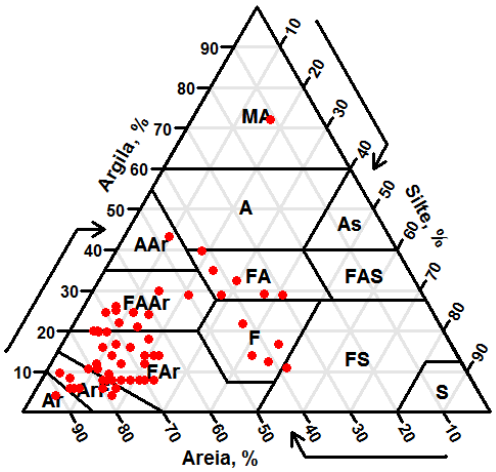


Geral (n = 401)

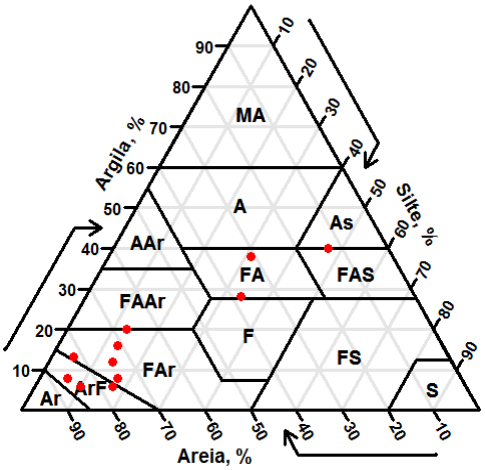
(b)



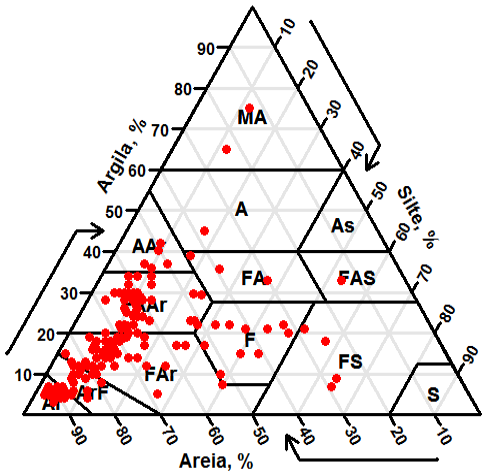
Geral (n = 70)



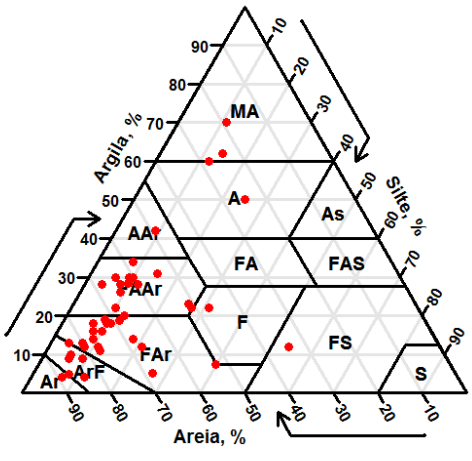
Argissolos (n = 64)



Argissolos (n = 11)



Latossolos (n = 337)
Desenvolvimento



Latossolos (n = 59)
Validação

Fonte: Próprio autor.

Ar: Areia; AF: Areia franca; FAr: Franco arenosa; FAAr: Franco argilo arenosa; AAr: Argila arenosa; A: Argila; MA: Muito argilosa; AS: Argila siltosa; FAS: Franco argilo siltosa; FS: Franco siltosa, S: Silte; F: franca; FA: Franco argilosa.

3.3.2 Desenvolvimento das funções de pedotransferência (FPTs)

Pelo método stepwise (direction = "both") as variáveis preditoras que influenciaram significativamente ($p<0,010$) a Ds foram inseridas no modelo, a fim de obter uma função de pedotransferência adequada às características dos solos em estudo.

As FPTs para o agrupamento de todas as classes de solos (Geral) e por classe estão apresentadas na Tabela 4. Para o desenvolvimento da FPT1 para todas as classes de solos, as variáveis selecionadas foram o teor de silte, densidade de partícula e o carbono orgânico do solo. Para o desenvolvimento da FPT2 para os Argissolos, foram consideradas no modelo o teor de silte, densidade de partícula e o carbono orgânico do solo. E para a FPT3 desenvolvida para os Latossolos os atributos inseridos no modelo foram o conteúdo de areia, conteúdo de silte, densidade de partículas e carbono orgânico do solo.

Tabela 6. Funções de pedotransferência (FPT) para todos os solos (Geral) e por classe de solo.

Modelo	FPT	Equação	n	R^2_{aj}
Geral	1	Ds = 0,35 – 0,006*Silte + 0,475*Dp – 0,061*CO	401	0,36
Argissolos	2	Ds = -0,26 – 0,004*Silte + 0,676*Dp – 0,043*CO	64	0,33
Latossolos	3	Ds = 0,22 + 0,002*Areia – 0,004*Silte + 0,465*Dp – 0,065*CO	337	0,34

Fonte: Próprio autor.

Os três modelos desenvolvidos (Tabela 4) apresentaram desempenho semelhante, sendo a FPT1 levemente melhor que os demais (R^2_{aj}>). Apesar da variável Dp não ser relatada como variável preditiva da Ds nos trabalhos disponíveis na literatura, possivelmente por este atributo não ser encontrado com frequência nos bancos de dados de solos, usá-lo não foge do propósito da função de pedotransferência (FPT) que, segundo Bouma (1989), é obter dados que precisamos por meio de dados que dispomos.

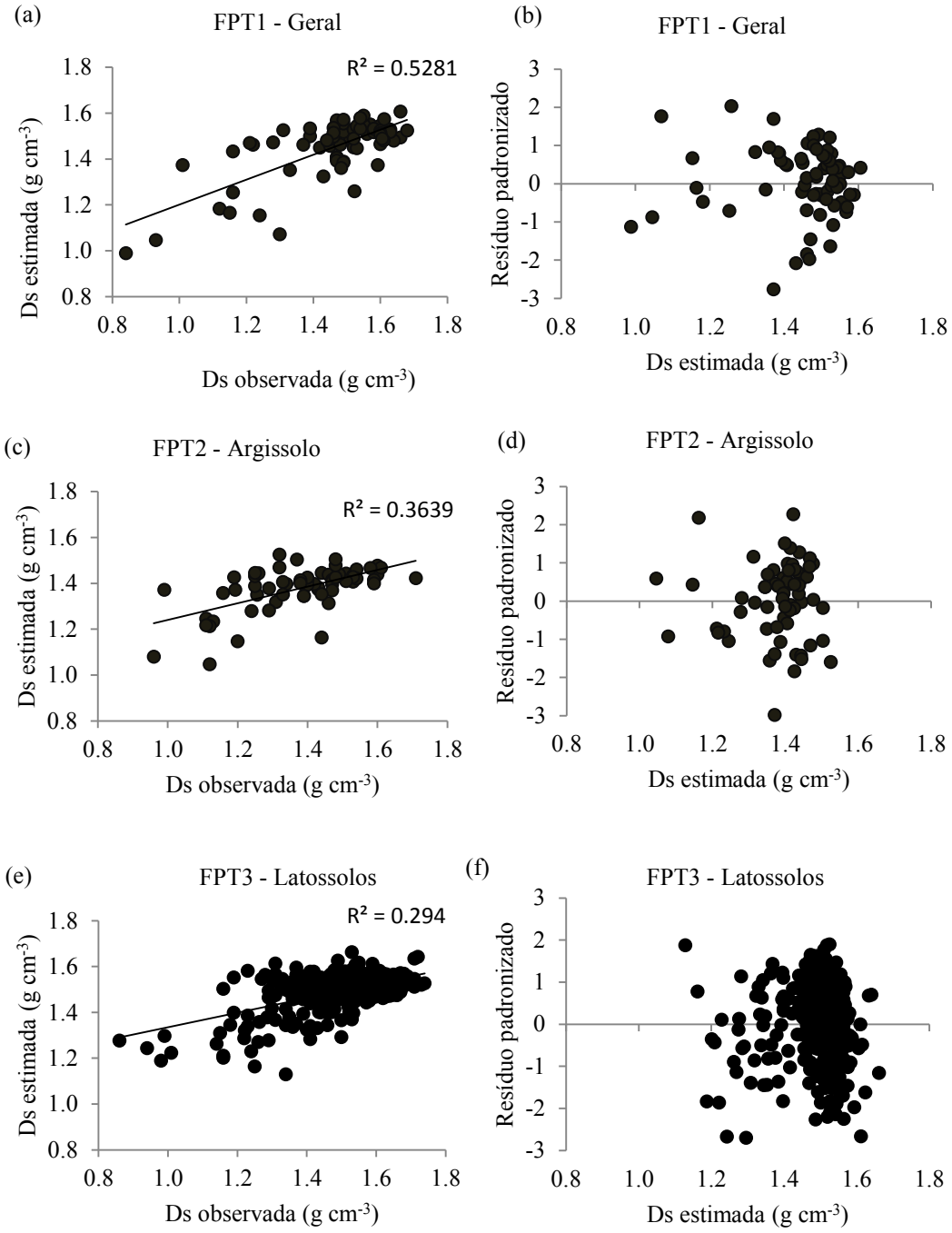
3.3.3 Avaliação das funções de pedotransferência (FPTs)

A acurácia das FPTs foi obtida por meio da comparação dos valores observados com os estimados (Figuras 3a, 3c e 3e), no mesmo conjunto de dados utilizado no desenvolvimento da função.

A comparação dos resíduos com os valores estimados para as FPTs geradas não apresentou qualquer padrão aparente, ou uma relação entre os resíduos e os valores estimados (Figuras 3b, 3d e 3f). Isso indica que os modelos ajustam os dados da melhor forma possível, dadas as variáveis independentes usadas na modelagem. Também não se observa valores

acima de +3 ou menores do que -3 que indica não haver dados discrepantes (outliers) nos bancos de dados utilizados

Figura 3 - Desempenho das funções de pedotransferência desenvolvidas e análise dos resíduos para estimativa da densidade do solo.



Fonte: Próprio autor.

Na Tabela 5 estão apresentados a avaliação do desempenho dos modelos de regressão, os indicadores estatísticos e o índice de desempenho. As FPTs desenvolvidas mostraram-se

ter bons ajustes para a densidade do solo. Sendo a de melhor desempenho a FPT desenvolvida para estimar os Argissolos. Pois esta apresentou valores do EM e RMSE mais próximos de zero, e valores de C e R² mais próximos e 1 mais apropriada é a FPT (CAMPELO JUNIOR et al. 2014).

Tabela 7 - Avaliação do desempenho dos modelos de regressão, os indicadores estatísticos e o índice de desempenho para todos os solos (Geral) e por classe de solo nos dados de validação.

Modelo	n	EM[13]	RMSE[14]	R^2	r	d[15]	C[16]	Desempenho
FPT1	70	0,01	0,12	0,50	0,71	1,00	0,71	Bom
FPT2	11	0,00	0,00	0,73	0,85	1,00	0,85	Muito Bom
FPT3	59	0,02	0,14	0,53	0,73	083	0,60	Mediano

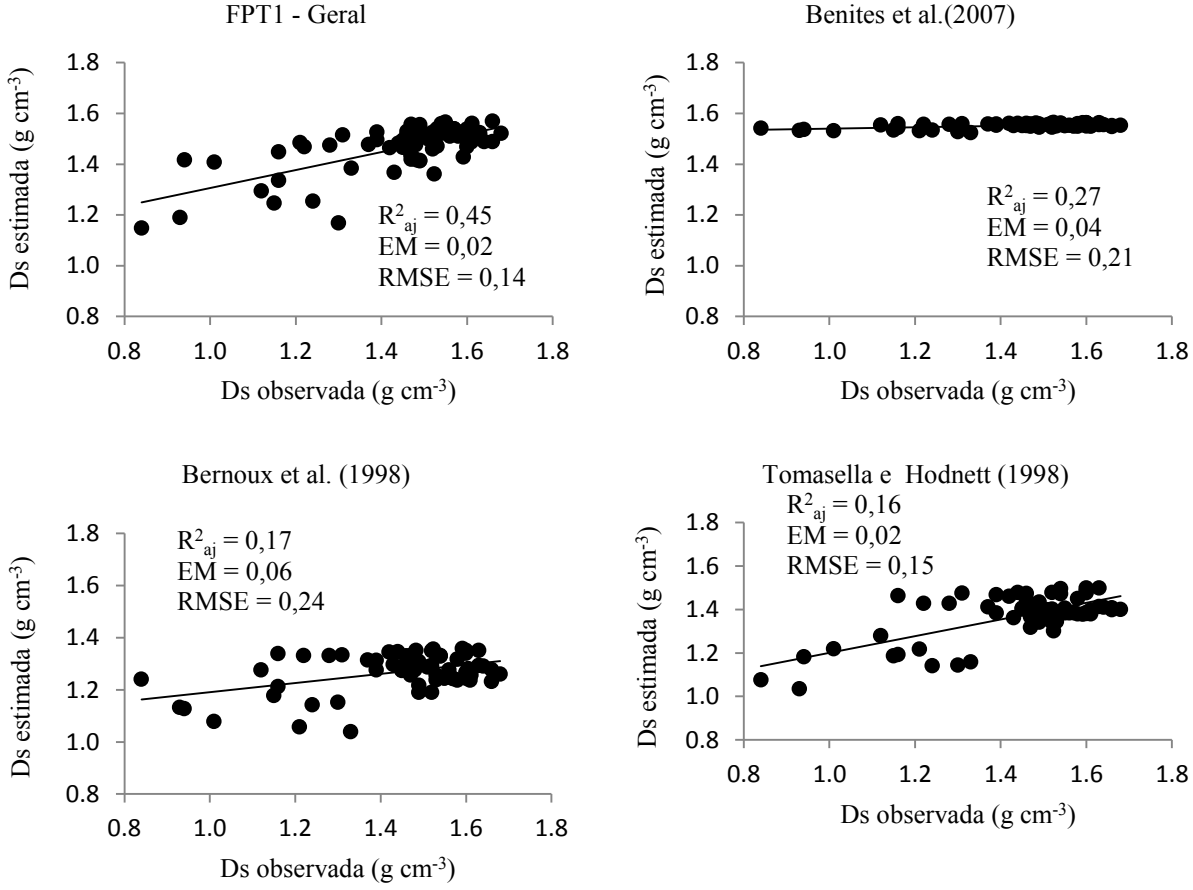
Fonte: Próprio autor.

3.3.4 Comparações com os modelos disponíveis na literatura

A análise da confiança das FPTs desenvolvidas foi verificada pela relação entre os valores observados são diferentes nos dados de validação (Figuras 4, 5 e 6). A FPT1 apresentou maior valor de R_{aj}^2 em comparação com as FPTs disponíveis na literatura, indicando maior confiança na utilização deste modelo. Os valores de R_{aj}^2 dos modelos disponíveis na literatura variaram de 0,16 a 0,27.

[13] EM = erro médio
[14] RMSE = raiz do quadrado do erro médio
[15] d = índice de Willmott
[16] C = Índice de desempenho

Figura 4 - Valores observados e estimados da função desenvolvida e das obtidas de outros estudos, considerando o conjunto de dados de validação (Geral). EM = erro médio; RMSE = raiz do quadrado do erro médio.



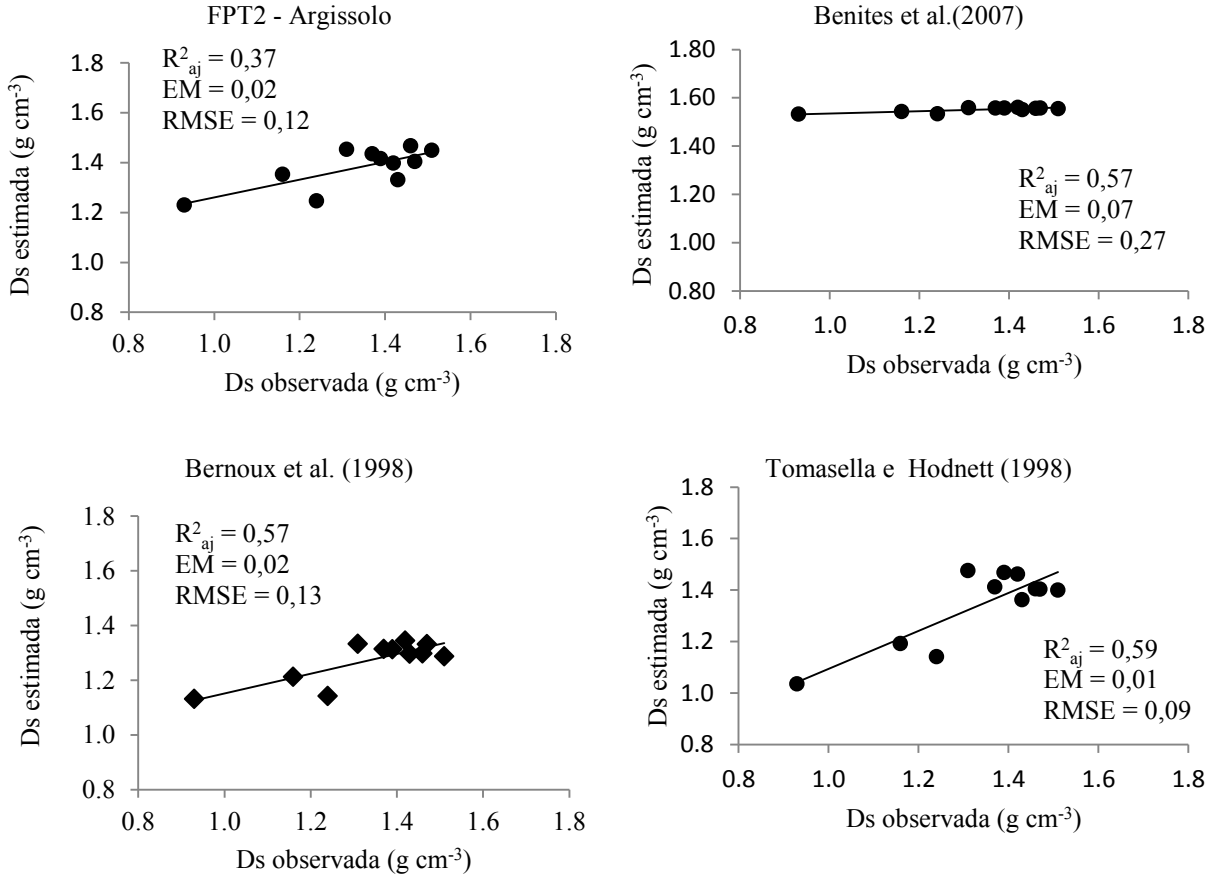
Fonte: Próprio autor.

Para os dados de Argissolos a FPT2 (Figura 5) desenvolvida a partir dos dados de validação apresentou baixa confiança em relação às funções disponíveis na literatura. Ao comparar a precisão nos dados de validação, a PTF desenvolvida por Tomasella e Hodnett (1998), utilizando solos da região Amazônica, mostrou melhor desempenho entre as funções já publicadas com R^2_{ajust} de 0,59, EM de 0,01 e RMSE de 0,09. Quanto menor forem o MSE e o RMSEP, melhor será a estimação da Ds em relação aos atributos do solo utilizados na construção do modelo.

Não houve grande diferença de desempenho das funções desenvolvidas neste estudo para a classe de Argissolos, comparada ao desempenho obtido por autores dos modelos disponíveis na literatura (Tabela 2) em seus solos de origem. Neste caso, é preferível a utilização de FPTs já validadas. A comparação das FPTs desenvolvidas com as disponíveis na literatura é importante na tentativa de utilização das funções já existentes. Diminuindo assim, o esforço na geração de novas funções, que de acordo com McBratney et al. (2002), embora

exista um grande número de FPT, são raras as iniciativas na tentativa de utilização das mesmas.

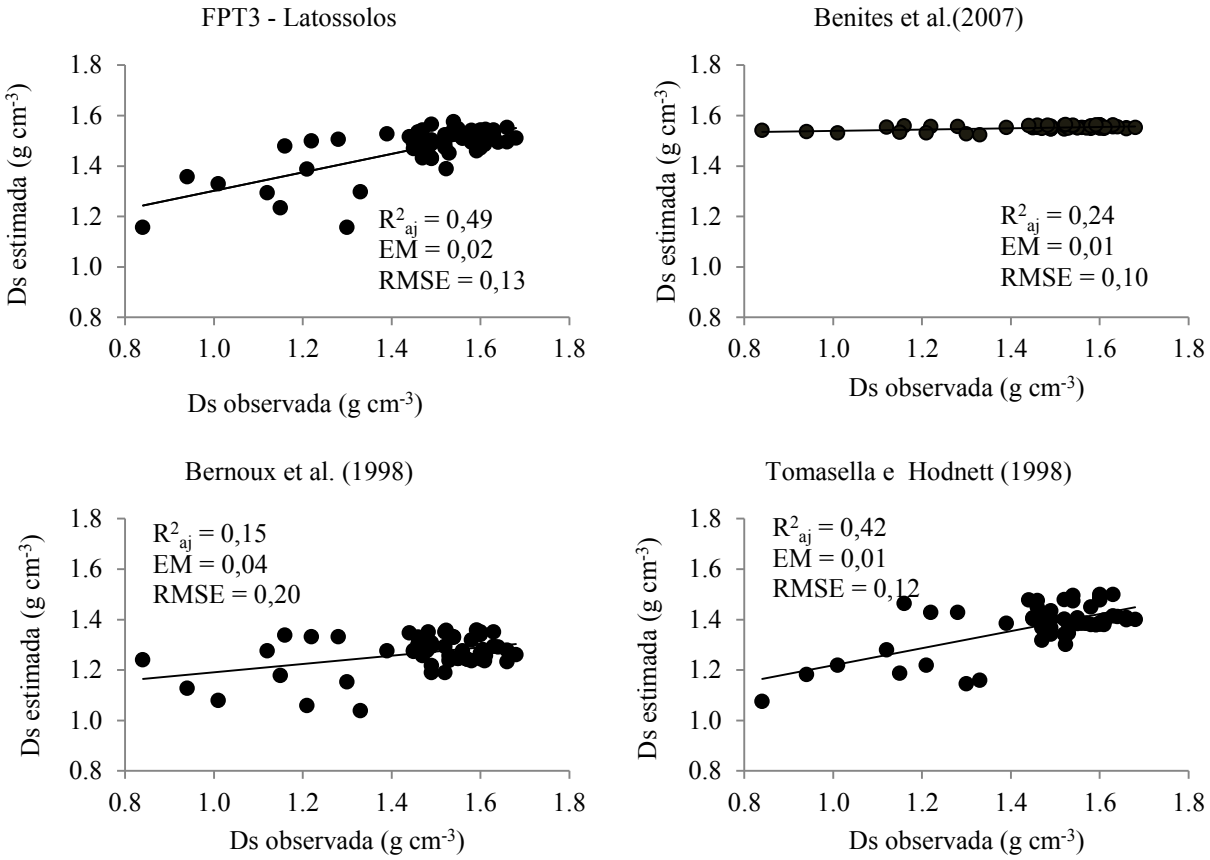
Figura 5 - Valores observados e estimados da função desenvolvida e das obtidas de outros estudos, considerando o conjunto de dados de validação (Argissolos). EM = erro médio; RMSE = raiz do quadrado do erro médio.



Fonte: Próprio autor.

No subconjunto dos dados de validação que compõe os Latossolos (FPT3), o modelo desenvolvido neste estudo, apresentou melhor ajustamento em relação aos demais modelos testados neste estudo.

Figura 6 - Valores observados e estimados da função desenvolvida e das obtidas de outros estudos, considerando o conjunto de dados de validação (Latossolos). MSE: erro quadrático médio; RMSEP: erro médio quadrático de previsão.



Fonte: Próprio autor.

Todos os modelos foram associados com valores positivos de erro médio (EM), variando de 0,01 a 0,07, indicando superestimativa da Ds. Quanto mais próximo de zero for o valor do EM e do RMSE, melhor será o desempenho da FPT (Pachepsky & Rawls, 2004). O EM igual zero significa que o estimador prevê observações com precisão perfeita.

Os resultados de EM e RMSE para o modelo proposto por Bernoux et al.(1998), assim como o gráfico de dispersão, mostraram que este modelo foi associado a maiores imprecisões e viés de previsão comparado aos demais modelos testados.

3.4 Conclusão

O modelo desenvolvido a partir do banco de dados Geral mostrou melhor desempenho que os modelos propostos pelos outros autores analisados neste estudo. A FPT desenvolvida por Tomasella e Hodnett (1998) dentre as analisadas, foi o que apresentou melhor desempenho.

REFERÊNCIAS

AKAIKE, H. Information theory and an extension of the maximum likelihood principle. , In: PETROV, B.N.; CSAKI, F. (editors). 2nd International Symposium on Information Theory. Budapest, Hungary: Akadémiai Kiadó, 1973. p. 267-281.

ALVARES, C.A.; STAPE, J.L.; SENTELHAS, P.C.; GONÇALVES, J.L.M.; SPAROVEK, G. Koppen's climate classification map for Brazil. Meteorologische Zeitschrift. Gebruder Borntraeger, Stuttgart, v.22, No. 6, p.711–728, 2013.

BARROS, H. S.; FEARNSIDE, P. M. Pedo-transfer functions for estimating soil bulk density in central amazonia, Revista Brasileira de Ciência do Solo, Viçosa, No.2, v.39, p.397-407, mar./apr. 2015.

BARTLETT, J.W.; MORRIS, T.P. Multiple imputation of covariates by substantive-model compatible fully conditional specification. Statistical Methods in Medical Research, v.24, No.4, p.462 – 487, aug.2015.

BERNOUX, M.; ARROUAYS, D.; CERRI, C.; VOLKOFF, B.; JOLIVET, C. Bulk densities of Brazilian Amazon soils related to other soil properties. Soil Science Socicfy of America Journal, v.62, No.3, p.743–749, may/jun. 1998.

BEUTLER, S. J.; PEREIRA, M. G.; TASSINARI, W. D. S.; MENEZES, M. D. D., VALLADARES, G. S., ANJOS, L. H. C. D., 2017. Bulk Density Prediction for Histosols and Soil Horizons with High Organic Matter Content, Revista Brasileira de Ciência do Solo, Viçosa, v.41, p.1-13, mar. 2017.

BENITES, V. M.; MACHADO, P. L.; FIDALGO, E. C.; COELHO, M. R.; MADARI, B. E. Pedotransfer functions for estimating soil bulk density from existing soil survey reports in Brazil, Geoderma, v. 139, No. 1, p. 90-97, feb. 2007.

BOUMA, J. Using soil survey data for quantitative land evaluation. In: STEWART, B.A., ed. Advances in soil science. New York, Springer Verlag,. v.9. p.177-213, 1989.

CAMARGO, A.P.; SENTELHAS, P.C. Avaliação do desempenho de diferentes métodos de estimativa da evapotranspiração potencial no Estado de São Paulo. Revista Brasileira de Agrometeorologia, Santa Maria, v. 5, p. 89-97, 1997.

CAMPELO JUNIOR, J. H.; AZEVEDO, E. C.; CARVALHO ALVES, M. de, MELLO, D. de; ALMEIDA LOBO, F. de; AMORIM, R. S. S. Estimativa da retenção de água em Latossolos do Cerrado mato-grossense cultivados com algodão. Revista agro@mbiente on-line, v. 8, n. 3, p. 318-326, 2014.

CARRERA, J.; NEUMAN, S.P. Estimation of Aquifer Parameters under Transient and Steady State Conditions I: Maximum Likelihood Method Incorporating Prior Information. Water Resources Research, v.22, p.199-210, feb. 1986.

DRAPER, N. R.; SMITH, H. Applied Regression Analysis. 3rd ed. New York: Wiley, 1998. 659 p.

CHEN, S.; FORGES, A. C. R.; SABY, N. P. A; MARTIN, M; WALTE, C; ARROUAYSA, D. Building a pedotransfer function for soil bulk density on regional dataset and testing its validity over a larger area. Geoderma, v.312, p.52–63, feb. 2018.

KHODAVERDILOO, H.; MOMTAZ H.; LIAO K. Performance of soil cation exchange capacity pedotransfer funmaction as affected by the inputs and database size. npj Clean Soil Air Water, v.46, No.3, p.1-8, mar. 2018.

LEMOS, R.C.; SANTOS, R.D. Manual de descrição e coleta de solo no campo. 3.ed. Campinas: Sociedade Brasileira de Ciência do Solo, 1996. 84p.

McBRATNEY, A.B.; MINASNY, B.; CATTLE, S.R.; VERVOORT, R.W. From pedotranfer functions to soil inference systems. **Geoderma**, Amsterdam, v. 109, p. 41-73, 2002.

MEDEIROS, J.C.; COOPER, M.; DALLA ROSA, J.; GRIMALDI, M.; COQUET, Y.,Assessment of pedotransfer functions for estimating soil water retention curves for the amazon region. Revista Brasileira de Ciência do Solo, Viçosa, v.38, No. 3, p.730–743, may/jun. 2014.

MINASNY, B.; MCBRATNEY, A.B.; BRISTOW, K.L. Comparison of different approaches to the development of pedotransfer functions for water-retention curves. Geoderma, v.93, p.225-253, dec.1999.

MYERS, R. H. Classical and Modern Regression with Applications., 2. ed. Boston: PWS – Kent Publishers, 1990. 488 p.

MOLDRUP, P.; OLESEN, T.; YOSHIKAWA, S.; KOMATSU, T.; ROLSTON, D. E. Three-porosity model for predicting the gas diffusion coefficient in undisturbed soil, Soil Science Society of America Journal, v.68, No.3, p.750–759, may. 2004.

NANKO, K.; UGAWA, S.; HASHIMOTO, S.; IMAYA, A.; KOBAYASHI, M.; SAKAI, H.; ISHIZUKA, S.; MIURA, S.; TANAKA, N.; TAKAHASHI, M.; KANEKO, S. A pedotransfer function for estimating bulk density of forest soil in Japan affected by volcanic ash. Geoderma, v.213, p.36-45, jan. 2014.

NEMES, A.; QUEBEDEAUX, B.; TIMLIN, D.J., Ensemble approach to provide uncertainty estimates of soil bulk density. Soil Science Society of America Journal. v.74, No.6, p.1938-1945, nov. 2010.

NGUYEN, P. M.; HAGHVERDI, A.; DE PUE, J.; BOTULA, Y. D.; LE, K. V.; WAEGEMAN, W.; CORNELIS, W. M. Comparison of statistical regression and data-mining techniques in estimating soil water retention of tropical delta soils. Biosystems Engineering, v.153, p. 12-27, jan. 2017.

NUNES, L.N.; KLÜCK, M.M.; FACHEL, J.M.G. Multiple imputations for missing data: a simulation with epidemiological data. Caderno de Saúde Pública, v.25, No. 2, p.268-278, feb. 2009.

PAES. Â.T., POLETO, F.Z. Por dentro da estatística. Educação Continuada em Saúde

einstein, v.11, p. 5–7, 2013.
PEÑA, E.A.; SLATE, E.H. Global Validation of Linear Model Assumptions. Journal of the American Statistical Association. v.101, p.341-354, mar. 2006.

PEÑA, E. A.; SLATE, E. H., 2019. gvlma: Global Validation of Linear Models Assumptions. R package version 1.0.0.3. Disponível em: http://CRAN.R-project.org/package=gvlma. (acessado em 09 de junho de 2019).

QIAO, J.; ZHU, Y.; JIA, X.; HUANG, L.; SHAO, M. Development of pedotransfer functions for predicting the bulk density in the critical zone on the Loess Plateau, China. Journal of Soils and Sediments, v. 19, p. 366 – 372, may. 2018.

R CORE TEAM., 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
ROUSSEAU, M.; SIMON, M.; BERTRAND, R.; HACHEY, K. Reporting missing data: a study of selected articles published from 2003-2007. Quality & Quantit,. v. 46, p. 1393–1406, aug. 2012.

RUBIN, D.B. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons, 1987. 258 p.

SANTOS, H.G. et al. Sistema Brasileiro de Classificação de Solos. 5. ed. Brasília, DF: EMBRAPA, 2018.

SCHAFER, J.L.; GRAHAM, J.W. Missing data: our view of the state of the art. Psychol Methods, v.7, No. 2, p.147-77, jun. 2002.

SEAMAN, S., HUGHES, R.A., 2016. Relative efficiency of joint-model and full-conditional-specification multiple imputation when conditional models are compatible: The general location model. Statistical Methods in Medical Research, v.27, No. 6, p. 1603–1614, sep. 2016.

SEQUEIRA, C.; WILLS, S.; SEYBOLD, C.; WEST, L. Predicting soil bulk density for incomplete databases. Geoderma, v.213, p.64–73, jan. 2014.

SNEE , R. D. Some aspects of nonorthogonal data analysis, Part I. Developing prediction equations. Journal of Quality Technology, v.5, p. 67 – 79, 1973.

STAVSETH, M.R.; CLAUSEN, T.; RØISLIEN, J. How handling missing data may impact conclusions: A comparison of six different imputation methods for categorical questionnaire data. SAGE Open Medicine, v.7, p.1–12, jan. 2019.

SUUSTER, E.; RITZ, C.; ROOSTALU, H.; REINTAM, E.; KÕLLI, R.; ASTOVER, A. Soil bulk density pedotransfer functions of the humus horizon in arable soils. Geoderma, v.163, p. 74–82, jun. 2011.

TOMASELLA, J.; HODNETT, M.G. Estimating soil water retention characteristics from limited data in Brazilian Amazonia. Soil Science, v.163, p.190–202, mar. 1998.

VAN BUUREN, S.; GROOTHUIS-OUDSHOORN, K. Mice: Multivariate Imputation by

Chained Equations in R. Journal of Statistical Software, v.45, p.1-67, dec. 2011.

WILLMOTT, C.J. On the validation of models. Physical Geography 2: p.184–194, 1981.

CONCLUSÕES GERAIS E RECOMENDAÇÕES

1. O algoritmo MICE se mostrou como uma boa alternativa para o tratamento de dados faltantes. As estimativas de dados de atributos físico-hídrico do solo, em geral, foram melhores do que as estimativas empregando o método convencional (Análise de casos completos). A baixa variabilidade entre os valores observados e os imputados indica uma boa precisão no processo de imputação;
2. Com a imputação de dados faltantes tornou-se possível manter variáveis pouco disponíveis (incompletas) em conjuntos de dados de solos, para serem utilizadas como variáveis preditoras no desenvolvimento de funções de pedotransferência (FPT);
3. Não podemos generalizar os resultados obtidos neste trabalho, outros cenários precisam ser estudados (diferentes tamanhos de amostras e proporção de dados faltantes). O presente estudo utilizou o método MICE para solucionar um problema existente em um conjunto de dados específico e para divulgar esta ferramenta.
4. Existem vários métodos para tratar dados faltantes. No entanto, é importante que se analise com cautela as características dos dados e os pressupostos de cada método, para garantir que a melhor ferramenta seja utilizada;
5. FPTs desenvolvidas para solos locais é preferível que o uso das FPTs disponíveis na literatura.

TRABALHOS FUTUROS

Para melhor avaliar os métodos propostos neste trabalho é interessante considerar:

- Simular as falhas no banco de dados com diferentes proporções de dados faltantes e comparar o método MICE com outros métodos;
- Comparar as funções de pedotransferência desenvolvida por meio do método de regressão linear múltipla com outros métodos para estimar a densidade do solo.

APÊNDICE A – ALGORITMO MICE NO R

Análise de casos completos e imputação de dados (Artigo 1)

 # Defining libraries

end_libs="~/Rpacks"

 # XLSX read package and its dependencies

library(readxl)

require(readxl)

suppressPackageStartupMessages(require(readxl,lib=end_libs))

 # Read Excel files and save to data frame

arq = read_excel("BD.xlsx","Plan2")

data<-arq

read<-data

 # Update

summary(data)

 # Listwise deletion of cases with missing values (Complete Case Analysis – CCA)

library(BaylorEdPsych)

summary(completeFC)

summary(lm(FC ~ Micro+Macro+PT+PMP+Ds+Argila+Silte+Areia+MO,data = arq))

lm(formula = FC ~ Micro+Macro+PT+PMP+Ds+Argila+Silte+Areia+MO,data = arq)

summary(lm(FC ~ Micro+Macro+Ds,data = arq))

lm(formula = FC ~ Micro+Macro+Ds,data = arq)

 # Quick classification of missing data

pMiss <- function(x){sum(is.na(x))/length(x)*100}

apply(data,2,pMiss)

apply(data,1,pMiss)

 # Little's MCAR test

library(mvnmle)

LittleMCAR(data)

 # Using mice for looking at missing data pattern

- install.packages(“mice”)

library(mice)

require(mice)

md.pattern(data)

Visual presentation of missing data pattern

- install.packages("VIM")

```
library(VIM)
require(VIM)
aggr_plot <- aggr(data, col=c('White','dark grey'), numbers=TRUE,prop=FALSE,
sortVars=TRUE, labels=names(data), cex.axis=.7, gap=3, ylab=c("Histogram of missing
data","Pattern"))
```

 # Imputing the missing data (MICE)

```
data.frame <- mice(data, m=5, maxit=10, meth='pmm')
summary(data.frame)
mice(data = data, m = 5, method = "pmm", maxit = 10, seed = 500)
imp<-mice(data, m=5, seed = 23109, print = FALSE)
fit<-with(imp, lm(FC ~ Micro+Macro+TP+PWP+Bd+Clay+Silt+Sand+OM))
summary(fit)
```

 # Analysis of imputed data (Pooling)

```
round(summary(pool(fit)), 3)
```

#Inspecting the distribution of original and imputed data

```
complete(imp)
plot(imp, 3:5)
densityplot(imp)
```

Reference

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

APÊNDICE B - ANÁLISE DESCRITIVA DO BANCO DE DADOS GERAL

Tabela 1. Quantitativo de amostras por município.

Municípios	Amostras
Baião	13
Barcarena	16
Belém	28
Bragança	142
Capitão Poço	220
Faro	1
Igarapé - Açú	43
Irituia	40
Itaituba	10
Juruti	3
Marabá-Carajás	2
Marapanim	16
Medicilândia	1
Mojú	4
Monte Alegre	1
Óbidos	1
Oriximiná	9
Paragominas	12
Prainha	7
Rurópolis	4
Santarém	2
São Miguel do Guamá	40
Tailândia	4
Uruará	12
Total	631

Fonte: Próprio autor.

Tabela 2. Distribuição de frequência das Classes de solos.

Tipo de solo	Amostras
Argissolo	226
Latossolo	398
Neossolo	7
Total Geral	631

Fonte: Próprio autor.

Tabela 3. Distribuição de frequência da classe textural dos solos.

Classe textural	Amostras
Areia fina	44
Areia fina franca	102
Argila	5
Argila arenosa	10
Franco	17
Franco arenoso	229
Franco argilo arenoso	189
Franco argilo siltoso	1
Franco argiloso	11
Franco siltoso	6
Franco-arenosa	3
Franco-argilo-arenosa	10
Muito Argilosa	4
Total	631

Fonte: Próprio autor.

Tabela 4. Percentuais de dados faltantes por variáveis.

Variáveis	Faltantes	
	n	%
PMP	171	27,10
MO	153	24,25
Micro	101	16,01
Macro	96	15,21
CC	86	13,63

Fonte: Próprio autor.

Tabela 5. Estatísticas decritivas das variáveis

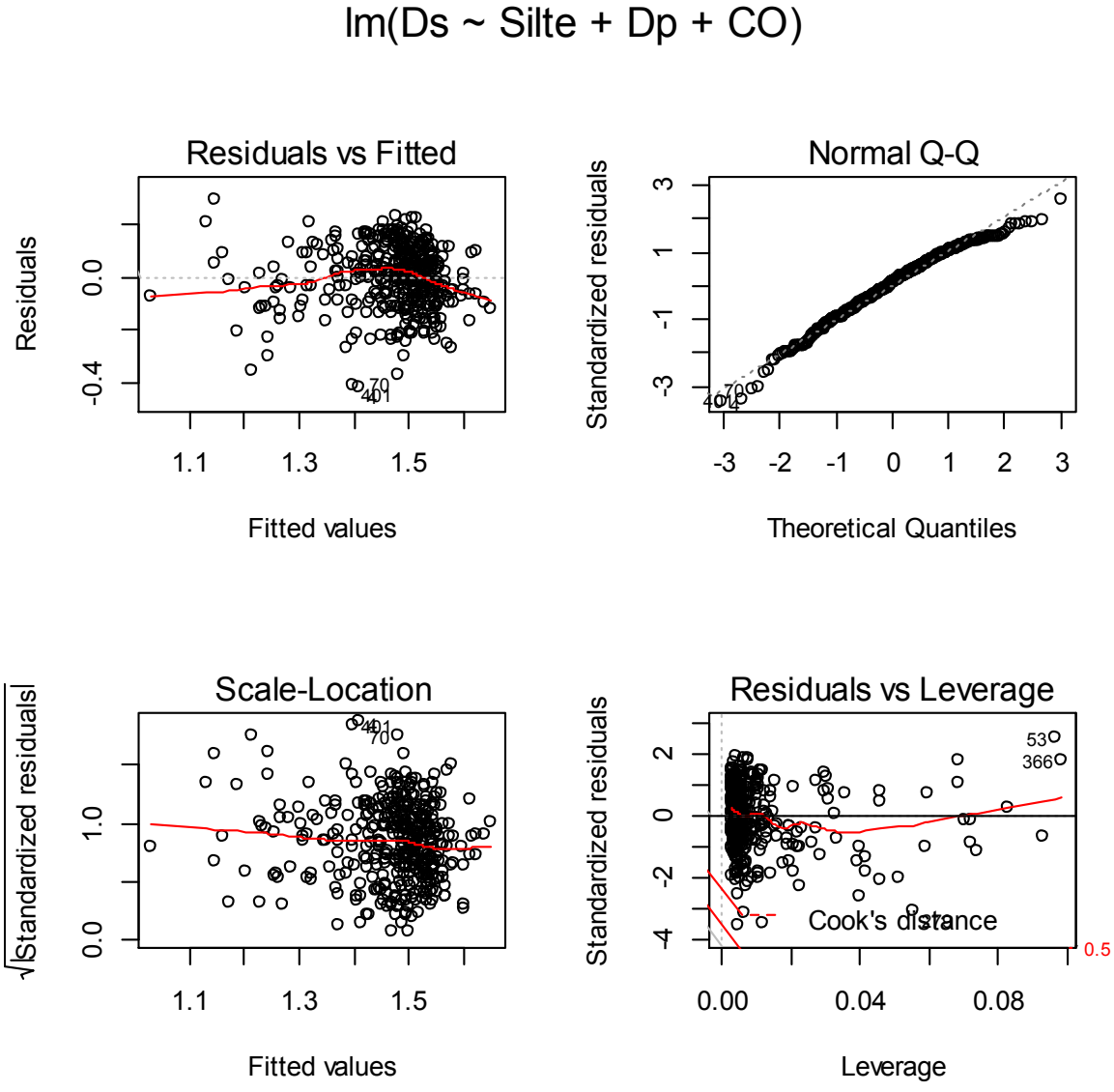
	Umidade (%)		Porosidade (%)			Densidade (g/cm3)		Textura (%)			MO
	CC	PMP	Total	Macro	Micro	Solo	Partículas	Areia	Silte	Argila	%
N	545	460	631	535	530	631	631	631	631	631	478
Média	22,45	10,63	44,19	21,29	22,34	1,45	2,61	70,36	11,93	17,87	1,39
Erro padrão	0,26	0,24	0,21	0,28	0,23	0,01	0,00	0,57	0,37	0,41	0,05
Mediana	22,45	9,64	44,00	21,00	22,46	1,47	2,61	73,00	9,00	16,00	0,99
Moda	25,70	7,33	46,00	16,70	25,70	1,47	2,60	74,00	6,00	14,00	1,13
Desvio padrão	5,98	5,24	5,36	6,45	5,19	0,15	0,07	14,20	9,36	10,23	1,19
CV	26,64	49,34	12,12	30,32	23,23	10,06	2,84	20,18	78,48	57,26	85,89
Curtose	1,62	4,85	1,77	-0,52	1,02	1,40	3,11	2,28	7,86	4,23	8,69
Assimetria	0,72	1,57	1,02	0,42	0,45	-0,94	-0,33	-1,24	2,63	1,38	2,58
Intervalo	36,60	35,60	37,00	29,29	34,50	0,90	0,67	84,00	63,00	72,00	7,72
Mínimo	8,50	2,30	31,00	10,19	8,50	0,84	2,22	11,00	1,00	3,00	0,18
Máximo	45,10	37,90	68,00	39,48	43,00	1,74	2,89	95,00	64,00	75,00	7,90
Nível de confiança (95,0%)	0,50	0,48	0,42	0,55	0,44	0,01	0,01	1,11	0,73	0,80	0,11

Fonte: Próprio autor.

APÊNDICE C - ANÁLISE DOS RESÍDUOS DAS FUNÇÕES DE PEDOTRANSFERÊNCIA DESENVOLVIDAS

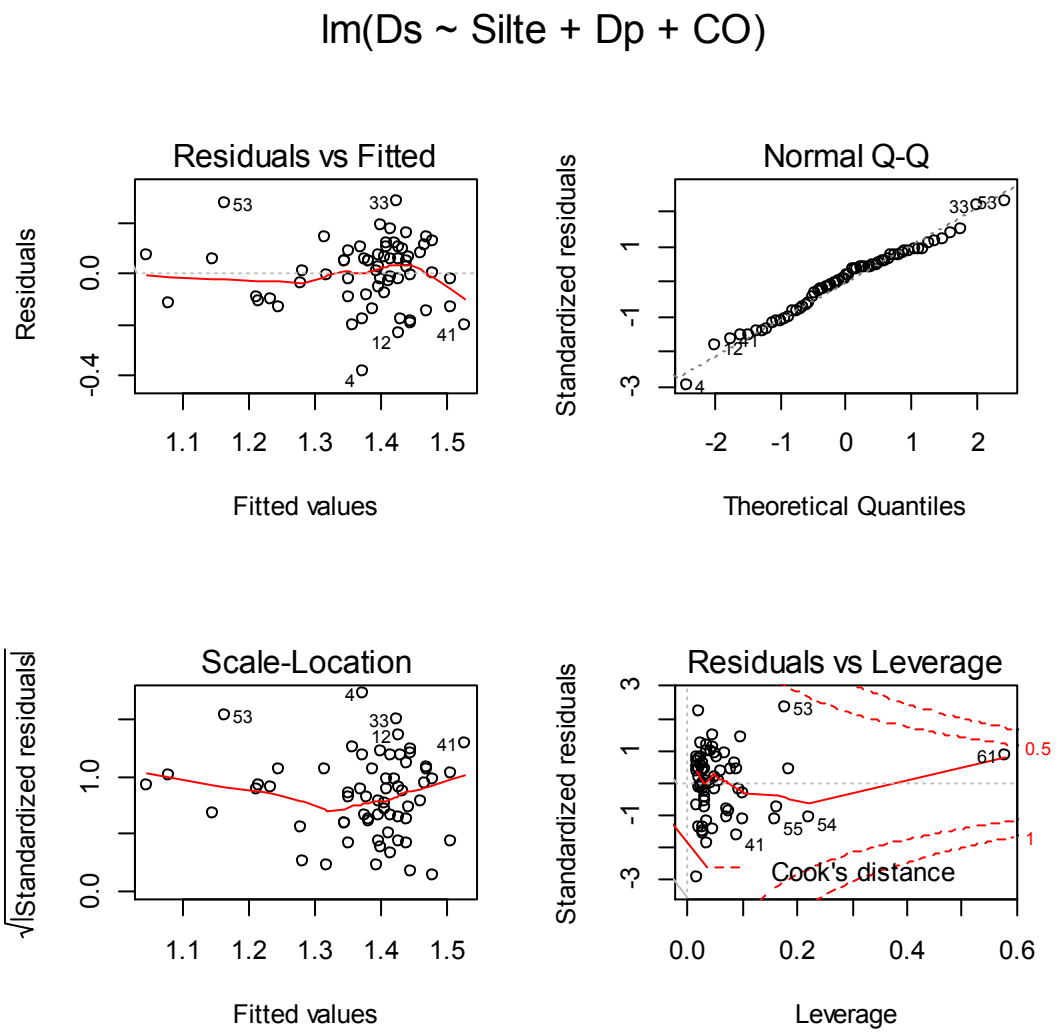
ARTIGO 2

Figura 1 - Análise dos resíduos da FPT1 (Geral)



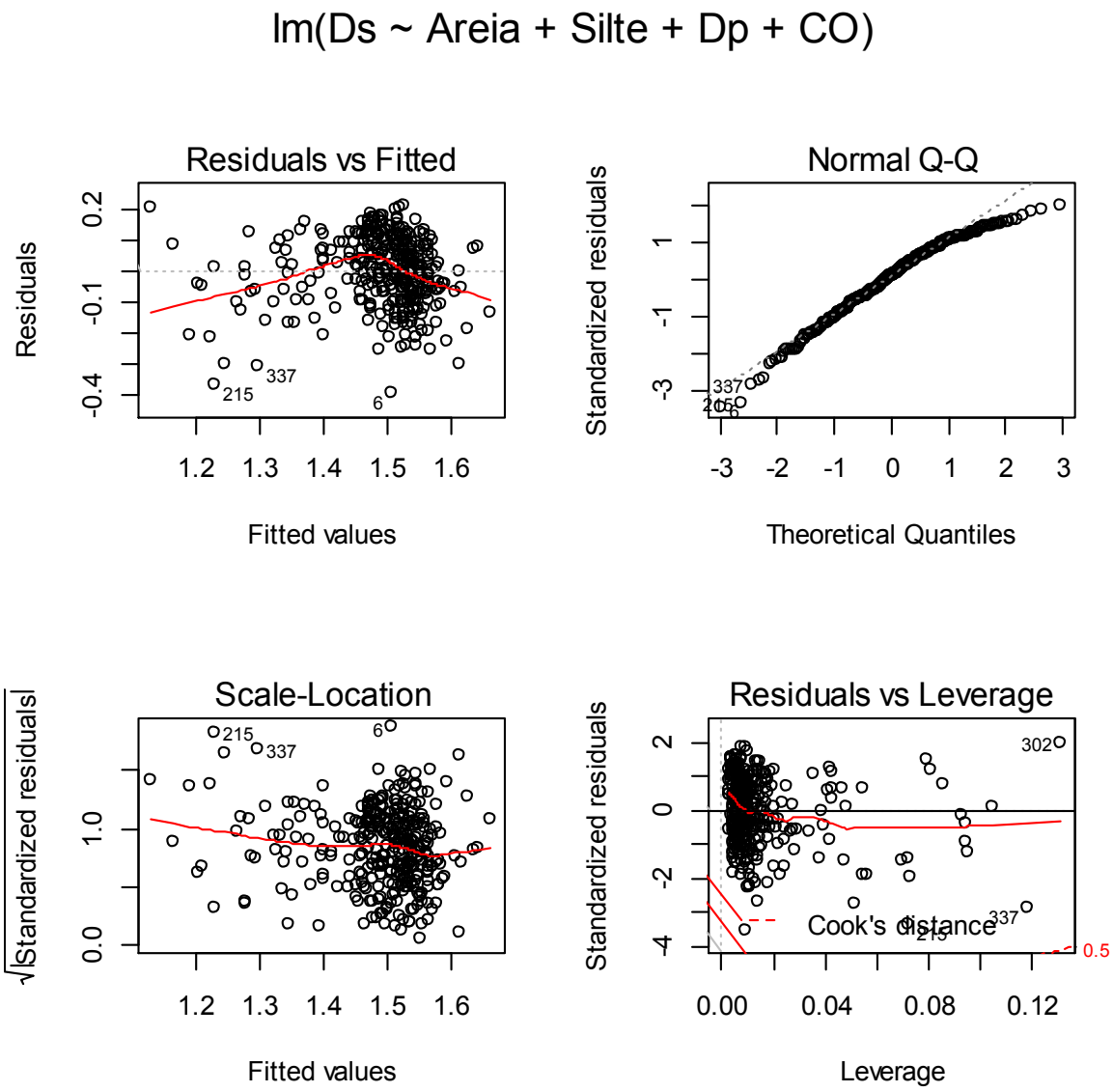
Fonte: Próprio autor.

Figura 2 - Análise dos resíduos da FPT2 (Argissolos)



Fonte: Próprio autor.

Figura 2 - Análise dos resíduos da FPT3 (Latossolos)



Fonte: Próprio autor.