UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE CIÊNCIAS RURAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DO SOLO

FUNÇÕES DE PREDIÇÃO ESPACIAL DE PROPRIEDADES DO SOLO

DISSERTAÇÃO DE MESTRADO

Alessandro Samuel Rosa

Santa Maria, RS, Brasil
2012

FUNÇÕES DE PREDIÇÃO ESPACIAL DE PROPRIEDADES DO SOLO

Alessandro Samuel Rosa

Dissertação apresentada ao Curso de Mestrado do Programa de Pós-Graduação em Ciência do Solo, Área de Concentração em Processos Físicos e Morfogenéticos do Solo, da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do grau de
Mestre em Ciência do Solo.

Orientador: Prof. Ricardo Simão Diniz Dalmolin

Santa Maria, RS, Brasil
2012

R788f Rosa, Alessandro Samuel
Funções de predição espacial de propriedades do solo / por Alessandro Samuel Rosa. – 2012.
201 p. ; il. ; 30 cm

Orientador: Ricardo Simão Diniz Dalmolin
Dissertação (mestrado) – Universidade Federal de Santa Maria, Centro de Ciências Rurais, Programa de Pós-Graduação em Ciência do Solo, RS, 2012

1. Mapeamento digital de solos 2. Atributos de terreno 3. Validação cruzada 4. Incerteza 5. Log-razão aditiva I. Dalmolin, Ricardo Simão Diniz II. Título.

CDU 631.4:528.7/.9

Ficha catalográfica elaborada por Cláudia Terezinha Branco Gallotti – CRB 10/1109
Biblioteca Central UFSM

Universidade Federal de Santa Maria
Centro de Ciências Rurais
Programa de Pós-Graduação em Ciência do Solo

A Comissão Examinadora, abaixo assinada,
aprova a Dissertação de Mestrado

FUNÇÕES DE PREDIÇÃO ESPACIAL DE PROPRIEDADES DO SOLO

elaborada por
Alessandro Samuel Rosa

como requisito parcial para obtenção do grau de
Mestre em Ciência do Solo

COMISSÃO EXAMINADORA:

Ricardo Simão Diniz Dalmolin, Dr.
(Presidente/Orientador)

Jean Paolo Gomes Minella, Dr. (UFSM)

Alexandre ten Caten, Dr. (IFF – Júlio de Castilhos)

Santa Maria, 27 de janeiro de 2012

A todo maluco que achar interessante “perder” seu tempo divagando,
dedico**.**

AGRADECIMENTOS

A serpente, que tentou Eva e Adão, fazendo-lhes provar o fruto proibido da árvore da ciência (do conhecimento do bem e do mal). Não fosse pela serpente nada teria sido possível:

- nem eu, meus pais, meu irmão e o resto da família, que, “bem ou mal”, contribuíram para minha formação;

- nem a Universidade Federal de Santa Maria, com seus professores, funcionários e alunos, que, “bem ou mal”, contribuíram para minha formação;

- nem a sociedade brasileira, financiadora de minha graduação e pós-graduação através do CNPq, que, “bem ou mal”, contribuíu para minha formação;

- nem o fantástico grupo de Pedologia, que, “bem ou mal”, contribuiu para minha formação;

- nem tudo o mais que existe, que, “bem ou mal”, contribuiu para minha formação.

A serpente, meu muito obrigado!

"[…] improvement in the prediction of soil properties does not rely on more sophisticated statistical methods, but rather on gathering more useful and higher quality data."
(Budiman Minasny e Alex McBratney, 2007)

APRESENTAÇÃO

Essa dissertação de mestrado é resultado da continuação do projeto de pesquisa do Setor de Pedologia do Departamento de Solos da Universidade Federal de Santa Maria intitulado *Avaliação da degradação ambiental e sustentabilidade de terras de encosta do Rebordo do Planalto – RS*. Dentro desse projeto de pesquisa guarda-chuva insere-se o projeto de pesquisa intitulado *Avaliação, modelagem e planejamento dos sistemas de uso da terra, e seu impacto sobre a qualidade do solo e da água, no Rebordo do Planalto do Estado do Rio Grande do Sul*, aprovado pelo edital MCT/CNPq/CT-Hidro nº 22/2009, do qual a presente dissertação de mestrado foi derivada.

Estruturei essa dissertação de mestrado em seis seções principais, a saber: 1) introdução, 2) revisão bibliográfica, 3) material e métodos, 4) resultados, 5) discussão, 6) conclusões, 7) Referências e, 8) Anexos. Entretanto, a ideia inicial era a de publicar a dissertação no formato de artigos científicos, e não no formato tradicional, conforme tem sido feito por diversos de meus colegas. Com o passar do tempo acabei ficando desconfortável com a ideia de publicação da dissertação no formato de artigos científicos, pois muita informação importante seria perdida, sobretudo no que diz respeito a seção Material e Métodos. A decisão definitiva de mudar a estrutura para o formato tradicional ocorreu apenas em novembro de 2011, depois de ouvir os comentários do Dr. Jean Paolo Gomes Minella sobre a estrutura de apresentação de trabalhos acadêmicos durante a prova de defesa da tese de doutorado do Dr. Alexandre ten Caten. Para ele um trabalho acadêmico deve ser completo, constando todos os detalhes necessários que permitam a reprodução do trabalho por outro pesquisador ou mesmo a sua verificação, ao contrário de um artigo científico que, pela limitação no número de páginas, geralmente contém apenas informações chave e omite detalhes fundamentais.

Apesar de, atualmente, concordar com o ponto de vista do Dr. Minella, sou sincero em dizer que a seção 2, referente à revisão bibliográfica, deixa a desejar no que diz respeito à fundamentação dos métodos estatísticos e matemáticos que utilizei. Isso porque o principal foco da revisão bibliográfica, que constitui a subseção 2.1, construída após o término de todas as demais seções e subseções da presente dissertação, é contribuir para a reflexão (nem que seja apenas a minha reflexão). Uma reflexão histórica crítica sobre o mapeamento de solos, tanto através do método tradicional como do "novo" (mapeamento digital), tentando compreender o cenário atual e trazer a baila elementos para discutirmos o futuro dessa

disciplina da ciência do solo no Brasil. A motivação para tentar fazer essa reflexão vem de leituras que tenho feito por iniciativa própria de autores como Edgar Morin, Paul Feyerabend, Bertrand Russell, entre outros. Além disso, foi decisivo um dos questionamentos a que fui submetido durante a realização da prova de seleção para o curso de doutorado em Agronomia-Ciência do Solo da Universidade Federal Rural do Rio de Janeiro, em novembro último: "qual o papel da ciência do solo frente a redução das desigualdades sociais?". Isso tudo me faz crer que precisamos refletir sobre o que fazemos, por que fazemos, como fazemos e para quem fazemos todos os dias dentro de nossos laboratórios com o dinheiro público. Raramente fui incentivado a fazer isso dentro da universidade. E quando o fiz recebi olhares reprovadores de muitos, cegados pelos seus dogmas. Provavelmente muitos entenderão de maneira errônea algumas das ideias que expresso a seguir, assim como o fizeram quando, ao falar em "falácia do plantio direto", entenderam que sou contrário ao sistema de plantio direto. Assim, antes que me crucifiquem, aviso que meu objetivo não é alcançar a "verdade", mesmo porque tal coisa não existe. Apenas tentei refletir e fazer refletir. Tentei construir meu próprio conhecimento e opinião sobre o mapeamento (digital) de solos no Brasil. E para isso lanço mão da linguagem polêmica, tão comum em outras áreas do conhecimento, mas que muitos ignorantes acabam confundindo com princípios gnosiológicos, confusão que para Gramsci só pode ser fruto da má-fé. Para aqueles que se sentirem incomodados, mesmo com essa apresentação, e não quiserem se aborrecer com meus devaneios, sugiro que saltem diretamente da introdução para a subseção 2.2. A partir daí talvez encontrem uma porção um pouco maior de racionalidade.

Quanto à comissão examinadora dessa dissertação de mestrado, a escolhi, em acordo com meu orientador, o Dr. Ricardo Simão Diniz Dalmolin, com o objetivo de poder ser, ao mesmo tempo, "alvejado" por dois profissionais que sabem muito mais do que eu sobre mapeamento digital de solos (o Dr. ten Caten) e sobre o uso e aplicação das informações contidas em mapas de solos (o Dr. Minella). Infelizmente não posso ter uma comissão examinadora composta, também, por representantes dos leigos que financiaram meus estudos. E nesse caso gostaria que fossem alguns dos moradores da bacia hidrográfica em que desenvolvi meu projeto (uma democracia direta, assim como ocorreu com as primeiras democracias da antiguidade, sobretudo na Grécia). Será que seria aprovado?

No que diz respeito às normas de apresentação de trabalhos acadêmicos da Universidade Federal de Santa Maria, assumo ter ignorado algumas. Registro isso porque discordo de várias delas, sobretudo aquelas que prejudicam a qualidade visual da apresentação do texto ou restringem a liberdade literária dos autores. Tais normas são, inclusive e

curiosamente, peculiares ao Brasil. Assim, optei por não separar os nomes de autores nas citações no texto por ponto e vírgula, mas sim utilizando a conjunção aditiva "e". Da mesma maneira, não redigi os nomes dos autores em letras maiúsculas, com exceção da primeira. Na lista de referência, redigi o nome de todos os autores, mesmo quando havendo mais de três, caso em que as normas da UFSM recomendam o uso da expressão et al. E, por último, utilizei tanto a voz ativa como a voz passiva, conforme julguei mais interessante e de acordo com o impacto que queria produzir no leitor. As normas da UFSM recomendam o uso do estilo impessoal, enquanto periódicos científicos como *Science* e *Nature* recomendam o uso da voz ativa pelo fato da mesma permitir a construção de textos mais claros, concisos e objetivos, o que creio ser fundamental. Mas, além disso, creio que o uso da voz ativa "aproxima" ainda mais o autor de sua obra, mostra que é o autor que está fazendo aquelas afirmações, mostra que o autor assume o que diz, não vindo a cair na comum falácia da imparcialidade e da neutralidade científica.

Por fim, espero contribuir, "bem ou mal[1]", para o desenvolvimento do mapeamento (digital) de solos no Brasil.

O Autor.

[1] "Na Genealogia da Moral, Nietzsche procura mostrar que os conceitos de bom e mau não são conceitos que se estabelecem de acordo com uma razão prática universal. Esses conceitos são expressões do modo de ser daqueles que avaliam. Quem avalia estabelece um valor, que, portanto, não é fato moral e sim uma interpretação moral." (Zatti, 2008).

RESUMO

Dissertação de Mestrado
Programa de Pós-Graduação em Ciência do Solo
Universidade Federal de Santa Maria

FUNÇÕES DE PREDIÇÃO ESPACIAL DE PROPRIEDADES DO SOLO

Autor: Alessandro Samuel Rosa
Orientador: Ricardo Simão Diniz Dalmolin
Data e local de defesa: Santa Maria, 27 de janeiro de 2012.

A possibilidade de mapear as propriedades dos solos através do uso de funções de predição espacial de solos (FPESe) é uma realidade. Mas seria possível construir FPESe para estimar propriedades como a distribuição do tamanho de partículas do solo (dtp) em um superfície geomorfológica jovem e instável, com elevada complexidade geológica e pedológica? O que seria considerado um bom desempenho nessas condições e que alternativas temos para melhorá-lo? Com esse trabalho tento encontrar respostas para essas questões. Para isso utilizei um conjunto de 339 amostras de solo de uma pequena bacia hidrográfica de encosta da região Central do RS. Modelos de regressão linear múltiplos foram construídos com atributos de terreno (elevação, índice de convergência, índice de potência de escoamento). As FPESe explicaram mais da metade da variância dos dados. Tal desempenho é semelhante àquele da abordagem tradicional de mapeamento de solos. Para algumas frações de tamanho o desempenho das FPESe pode chegar a 70%. As maiores incertezas ocorrem nas áreas de maior heterogeneidade geológica. Assim, melhorias significativas nas predições somente poderão ser alcançadas se dados geológicos acurados forem disponibilizados. Enquanto isso, FPESe construídas a partir de atributos de terreno são eficientes em estimar a dtp de solos de regiões com geologia complexa e elevada instabilidade. Mas restam dúvidas que não consegui resolver! O mapeamento de solos é importante para a resolução dos principais problemas sociais e ambientais do nosso tempo? E se nossas atividades estivessem submetidas ao controle da população como em uma democracia direta, seriam elas dignas de receber atenção?

Palavras-chave: mapeamento digital de solos, atributos de terreno, incerteza, validação cruzada, log-razão aditiva.

ABSTRACT

Master's thesis
Graduation Program in Soil Science
Federal University of Santa Maria

SPATIAL PREDICTION FUNCTIONS OF SOIL PROPERTIES

Author: Alessandro Samuel Rosa
Advisor: Ricardo Simão Diniz Dalmolin
Place and date of the defense: Santa Maria, January 27, 2012.

The possibility of mapping soil properties using soil spatial prediction functions (SSPFe) is a reality. But is it possible to SSPFe to estimate soil properties such as the particle-size distribution (psd) in a young, unstable and geologically complex geomorphologic surface? What would be considered a good performance in such situation and what alternatives do we have to improve it? With the present study I try to find answers to such questions. To do so I used a set of 339 soil samples from a small catchment of the hillslope areas of central Rio Grande do Sul. Multiple linear regression models were built using land-surface parameters (elevation, convergence index, stream power index). The SSPFe explained more than half of data variance. Such performance is similar to that of the conventional soil mapping approach. For some size-fractions the SSPFe performance can reach 70%. Largest uncertainties are observed in areas of larger geological heterogeneity. Therefore, significant improvements in the predictions can only be achieved if accurate geological data is made available. Meanwhile, SSPFe built on land-surface parameters are efficient in estimating the psd of the soils in regions of complex geology. However, there still are questions that I couldn't answer! Is soil mapping important to solve the main social and environmental issues of our time? What if our activities were subjected to a social control as in a direct democracy, would they be worthy of receiving any attention?

Key words: digital soil mapping, land-surface parameters, uncertainty, cross-validation, additive log-ratio.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE EQUAÇÕES

LISTA DE ANEXOS

SUMÁRIO

1 INTRODUÇÃO

Conhecer e compreender o ambiente em que estamos inseridos! Essa é a principal busca da humanidade, fundamental para que possamos desenvolver meios de modificá-lo e assim permitir a reprodução de nossa vida. Ao longo dos milênios essa busca levou ao desenvolvimento do gigantesco corpo de conhecimentos hoje disponível. Mas ainda estamos muitíssimo longe de compreender o ambiente que nos rodeia em sua totalidade. Sobretudo se estivermos tratando de componentes ambientais complexos como o é o caso do solo, o qual está diretamente relacionado às outras esferas que compõe os ecossistemas terrestres (Figura 1).

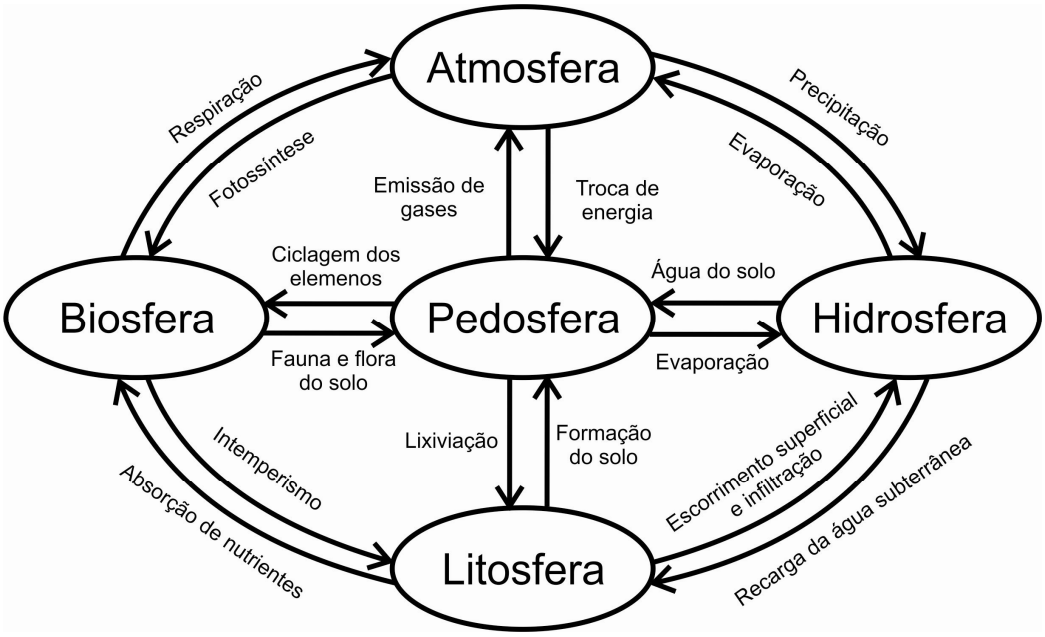


Figura 1 – Processos interativos ligando a pedosfera a litosfera, biosfera, hidrosfera e atmosfera. Adaptado de Lal et al. (1997).

Devido à intricada relação com a litosfera, a biosfera, a atmosfera e a hidrosfera, o solo é considerado por muitos como o principal componente na manutenção da qualidade ambiental. A maneira como o solo irá desempenhar essa função ambiental depende de suas propriedades. Dentre essas, a mais importante é a distribuição do tamanho de partículas, uma vez que ela determina e influencia a maioria das demais propriedades e atributos do solo. Isso

implica na necessidade de conhecermos a distribuição do tamanho de partículas dos solos de todo o planeta Terra (e inclusive de outros planetas se desejarmos colonizá-los).

Se fossemos utilizar a abordagem tradicional de mapeamento de solos, baseado no modelo *clorpt* de Hans Jenny, delineando mapas cloropléticos de solos, conhecer a distribuição do tamanho de partículas dos solos de todo o planeta Terra em escala elevada (> 1:25.000) dependeria da amostragem e análise de um conjunto extremamente grande de dados de solos. Creio que o número de amostras de solo necessário para cumprir essa tarefa é muitas vezes maior do que o número de amostras de solo já analisadas em todo o mundo desde o nascimento da ciência do solo (será que alguém possui uma estimativa desse número?). É óbvio que os esforços necessários para isso são muito grandes e possuem um elevado custo associado, tornando a tarefa preaticamente impossível de ser realizada dentro de uma ou duas décadas.

Mas o desenvolvimento e maior disponibilidade de computadores, bem como pacotes estatísticos e o sistema de informações geográficas, abriu uma nova possibilidade: construir modelos matemáticos para estimar as propriedades do solo em função de outras variáveis ambientais (funções de predição espacial de solos). Trata-se do mapeamento preditivo ou mapeamento digital de solos. Tal abordagem não enfrenta a maioria dos ditos "problemas" do procedimento tradicional de mapeamento de solos. Contudo, por se tratar de uma abordagem nova, muitas perguntas ainda estão sem resposta. Talvez a principal delas se refira ao fato de não sabermos se toda e qualquer superfície geomorfológica pode ter os solos mapeados através da abordagem preditiva (não estou comparando-a com o procedimento tradicional de mapeamento de solos). O fator que estabeleço como condicionante, nesse caso, é a necessidade de que as funções de predição espacial de solos tenham um desempenho tal que permita seu uso, com segurança, para avaliar ameaças ao solo e suas funções ambientais. Como seria esse desempenho em superfícies geomorfológicas jovens e instáveis, com elevada complexidade geológica e pedológica, como é o caso do Rebordo do Planalto do Rio Grande do Sul? O que seria considerado um bom desempenho nessas condições? Que alternativas temos para melhorar esse desempenho? Nas páginas seguintes tento responder essas questões e, principalmente, criar outras.

2 REVISÃO BIBLIOGRÁFICA

> "O conhecimento fragmentado produzido pela ciência é cada vez mais destinado a não ser meditado, refletido e discutido por espíritos humanos [...]. A ciência continua cega em relação a si mesma. Ela não compreende nem as causas nem as consequências de sua ação". (Edgar Morin, 1986)

2.1 Mapeamento (digital) de solo no Brasil – de onde viemos, onde estamos e para onde vamos

2.1.1 Nascimento, apogeu, morte e ressurreição

O ser humano sempre esteve, permanentemente, rodeado de quantidades maciças de *dados* que, quando capturados através dos órgãos sensoriais e processados por seu aparato cognitivo, mostram um mundo preenchido por *informação*. O entendimento desse mundo preenchido por informação depende, sobretudo, da capacidade do ser humano em classificar, capacidade essa desenvolvida em sua busca para encontrar ordem no universo (Krasilnikov et al., 2009), mesmo que a *ordem* não constitua a realidade (Gleiser, 2010). Assim, a necessidade de classificar é inerente ao desenvolvimento do conhecimento humano bem como fundamental para permitir a sua sobrevivência. A partir dessa necessidade e capacidade de classificar, o ser humano criou ao longo do tempo o que se chama áreas de conhecimento, definidas atualmente como o "conjunto de conhecimentos inter-relacionados, coletivamente construído, reunido segundo a natureza do objeto de investigação com finalidades de ensino, pesquisa e aplicações práticas" (CAPES, 2011). Uma dessas áreas do conhecimento é a bem conhecida *ciência do solo*, ocupada com o estudo do recurso natural solo, o que inclui aspectos relacionados à sua formação, classificação, mapeamento, suas propriedades e a

relação dessas com o uso e manejo do solo (Wikipédia, 2011).

A observação de que as propriedades do solo diferem de local para local, em função dos processos e fatores de formação, condicionando usos e manejos diferenciados, levou a criação de diversos *sistemas* de classificação de solos (Krasilnikov e Arnold, 2009). O conhecimento produzido a partir desses sistemas de classificação, e dos estudos deles derivados, foi fundamental para permitir o desenvolvimento da agricultura e o aumento da produção de alimentos no último século. Devido à importância do conhecimento pedológico, muitos países investiram em pesquisas na área, sobretudo incentivando a realização de levantamentos de solos. No Brasil os primeiros levantamentos de solos foram realizados na década de 1950, a partir da institucionalização da Comissão de Solos junto ao Ministério da Agricultura em 1947 (Mendonça-Santos e Santos, 2006). Mas foi a partir de 1970, com os projetos Radam e RadamBrasil, e forte investimento do governo federal, que os levantamentos de solos no Brasil ganharam força, a maioria em nível exploratório e de reconhecimento (Mendonça-Santos e Santos, 2006). Em meados da década de 1980 a evolução do conhecimento pedológico brasileiro atingiu seu auge (Ramos, 2003), sobretudo com a compilação do Mapa de Solos do Brasil, publicado na escala de 1:5.000.000, considerado até então a maior contribuição em termos de conhecimento de solos tropicais e subtropicais do mundo (Mendonça-Santos e Santos, 2006).

Contudo, a partir do final da década de 1980 o governo brasileiro começou a cortar, gradativamente, os incentivos para realização de levantamentos sistemáticos de solos (Ramos, 2003), assim como já havia ocorrido na Austrália na década de 1970, vinha ocorrendo na Nova Zelândia e Inglaterra e viria a ocorrer nos Estados Unidos na década de 1990 (Basher, 1997; Hartemink e McBratney, 2008; Finke, 2012). Isso levou ao abandono quase que total das atividades de levantamento de solos, ficando restritas às iniciativas individuais de instituições de ensino e pesquisa ou de empresas privadas. Muitos pesquisadores se puseram a refletir sobre os principais motivos que levaram a redução dos financiamentos e ao desmantelamento de muitos grupos de trabalho em todo o mundo. Para Basher (1997) um dos motivos foi a mudança do tipo de informações requeridas pela sociedade, que passou do qualitativo para o quantitativo. Segundo o autor os levantamentos de solos tradicionais são incapazes de responder a essa demanda com qualidade, baixo custo e em curto prazo. Essa incapacidade estaria relacionada a uma série de características dos mapas de solos obtidos através desse procedimento (Basher, 1997; Ker e Novais, 2003; Mendonça-Santos e Santos, 2003; Hartemink e McBratney, 2008), como a perda parcial de informação sobre variabilidade, a ênfase excessiva na classificação taxonômica, a utilização de terminologia

especializada, entre outras. Tais características se devem, sobretudo, ao fato de o modelo tradicional de mapeamento de solos, chamado discreto, materializado em mapas cloropléticos (ten Caten et al., 2011d), impor limites abruptos "arbitrários" entre as classes de solo definidas. Lembre que o domínio da ciência do solo (o solo) é contínuo no espaço, diferente da maioria das classificações onde os indivíduos classificados são discretos (Krasilnikov et al., 2009). Com essa imposição de limites abruptos a variação horizontal interna das unidades de mapeamento é desconsiderada e o valor das propriedades do solo nos locais não amostrados é estimado pelo valor médio da unidade de mapeamento, o que Mendonça-Santos e Santos (2003) indicam como sendo uma incapacidade de representar a realidade. Além disso, ao pautarem-se pela expressão de informações qualitativas, os mapas de solos (cloropléticos) obtidos através do modelo tradicional de mapeamento, não seriam capazes de expressar a complexidade dos solos na paisagem de uma maneira fácil de ser entendida (Sanchez et al., 2009). Isso estaria dificultando a aplicação prática da informação contida nos levantamentos pedológicos (Ker e Novais, 2003). Alguns autores chegam a afirmar que ao ignorar a variabilidade espacial dos solos o paradigma do modelo tradicional de mapeamento de solos é cientificamente inadequado (Mendonça-Santos e Santos, 2003).

No caso específico do Brasil, Ramos (2003) afirma que os principais motivos que levaram a redução dos incentivos governamentais para levantamentos de solos a partir da década de 1980 foram a crise orçamentária e financeira do país, bem como o fato de a maioria dos tomadores de decisão desconhecerem a importância dos levantamentos de solos para o planejamento de uso da terra. Ker e Novais (2003) afirmam que também pesou o fato de a pesquisa aplicada passar a ser privilegiada em relação à básica, assim como já havia sido sugerido por Basher (1997). Já Hartemink e McBratney (2008) e Grunwald (2009) atestam que um dos problemas mais importantes da ciência do solo mundial (e principalmente da pedologia) é que esteve sempre ligada muito diretamente à agricultura e, de uma maneira geral, olhando para dentro de si mesma. Nesse cenário houve falta de comunicação entre os pesquisadores, agricultores e técnicos (Ker e Novais, 2003). Com isso não foi possível produzir respostas quantitativas a questões antigas de uma maneira que pudesse ser utilizada diretamente por aqueles que usam o solo ou os tomadores de decisão (Hartemink e McBratney, 2008). Demorou muito para que os cientistas do solo de todo o mundo começassem a voltar sua atenção para outras questões que não o solo agrícola (pecuário e florestal), como aquelas relacionadas à mudança climática, regulação ambiental e serviços ambientais (Hartemink e McBratney, 2008), dando origem à expressão "*environmental soil science*" (Hillel, 2005). Como consequência, durante muito tempo e para muitos tomadores de

decisão, a função exclusiva do pedólogo era apenas a de mapeador do solo (Ramos, 2003). Talvez por isso, no Brasil, ao completar-se a Carta de Solos do Brasil na década de 1980, tenha-se entendido que não era mais necessária a função do pedólogo: ele já havia cumprido sua função (Ker, 1999; Ramos, 2003). O mesmo fenômeno parece ter acontecido em outros países (Finke, 2012). Com a redução da demanda governamental por informações pedológicas, parte significativa dos estudos pedológicos e levantamentos de solos realizados a partir de então no Brasil foi dirigida, principalmente, pela capacidade criativa e preocupações dos próprios pedólogos.

Mas na última década a pedologia (e a ciência do solo como um todo) conheceu um período de renascença (Hartemink e McBratney, 2008) que dura até os dias atuais e, aparentemente, deverá durar mais algum tempo. Isso ocorreu porque a demanda mundial por informações de solos cresceu de maneira acelerada (ten Caten, 2011), sobretudo pela necessidade de solucionar alguns dos ditos *principais problemas de nosso tempo*: segurança alimentar, mudança climática, degradação ambiental, escassez de água e as ameaças à biodiversidade (Ramos, 2003; Sanchez et al., 2009). Legislações foram criadas em diversos países para regular o uso do solo, o Programa de Desenvolvimento das Nações Unidas passou a dar cada vez mais ênfase ao solo em seus relatórios anuais, e o Painel Intergovernamental para a Mudança Climática reconheceu a importância do solo na mitigação dos gases de efeito estufa (Hartemink e McBratney, 2008). Mas para ser possível atender à nova demanda foi preciso mudar o foco das pesquisas (e o paradigma científico), conforme já havia sido sugerido por diversos pesquisadores (Basher, 1997; Ker e Novais, 2003; Ramos, 2003). A motivação passou a ser a de coletar e produzir dados de acordo com as necessidades específicas dos usuários da informação (*user-driven* ou *demand-driven*), dando menos ênfase a classificação taxonômica e mais ênfase a interpretação dos dados para o entendimento e modelagem de processos que ocorrem no solo (Basher, 1997). Não basta mais classificar e inventariar, mas sim compreender e quantificar, espaço-temporalmente, os padrões do solo em relação aos ciclos hidrológicos e a qualidade dos ecossistemas (Grunwald, 2009). É preciso que os cientistas do solo produzam mapas de alta resolução das propriedades funcionais do solo que são relevantes para o usuário (Sanchez et al., 2009). Para isso é necessário lançar mão de novas técnicas de coleta e processamento de dados e desenvolvimento de modelos que expressem a relação entre o solo e o seu ambiente de ocorrência, o que constitui o escopo do mapeamento digital de solos.

2.1.2 O mapeamento digital de solos

O mapeamento digital de solos (MDS) é definido como “a criação e população de sistemas de informação espacial de solos através do uso de métodos observacionais de campo e laboratório, acoplados a sistemas de inferência espacial e não-espacial de solos” (Lagacherie e McBratney, 2006). Utiliza-se, para isso, da pedometria, disciplina da ciência do solo correspondente a “aplicação de métodos matemáticos e estatísticos para o estudo da distribuição e gênese dos solos” (Heuvelink, 2003), dando origem às chamadas funções de predição espacial de solos com erros autocorrelacionados espacialmente (FPESe).

As bases do MDS foram estabelecidas formalmente por Alex McBratney, Maria de Lourdes Mendonça-Santos e Budiman Minasny em 2003, apesar de seu conceito ser utilizado em estudos pedológicos há bastante tempo (Troeh, 1964). Mas a consolidação do MDS em meio à comunidade científica internacional se deu somente a partir da criação do grupo de trabalho em MDS na União Internacional de Ciência do Solo em 2004, cuja iniciativa constituiu o consórcio GlobaSoilMap.net http://www.globalsoilmap.net/em 2006. Um dos objetivos do consórcio é a produção de mapas de propriedades funcionais do solo com cobertura global publicados em diversas resoluções, cada uma definida de acordo com as necessidades específicas dos usuários da informação (Sanchez et al., 2009). Dentre as propriedades do solo consideradas fundamentais pode-se citar o teor de carbono orgânico, a distribuição do tamanho de partículas, o pH, a capacidade de troca de cátions, a condutividade elétrica e a densidade (Carré et al., 2007; Sanchez et al., 2009). Os produtos do MDS podem ser utilizados para a modelagem quantitativa de propriedades do solo de difícil mensuração (Carré et al., 2007), podendo-se usar para isso funções de pedotransferência (McBratney et al., 2002; Sanchez et al., 2009). Tais propriedades, como a capacidade de retenção de água, estoque de nutrientes, depleção de nutrientes, suscetibilidade a erosão, entre outras, podem ser utilizadas para avaliar ameaças ao solo e suas funções, permitindo a simulação de cenários relacionados ao solo para a orientação da formulação de políticas públicas (Carré et al., 2007). Esse enfoque no mapeamento de propriedades do solo, levando em consideração a sua variação espacial, viria responder às necessidades atuais para as quais o mapeamento de solos através do método tradicional não fornece uma solução satisfatória.

Devido a essa mudança de paradigma no mapeamento de solos (McBratney et al., 2003), a maior parte dos trabalhos desenvolvidos nos últimos anos foca no mapeamento digital de propriedades do solo (McBratney et al., 2003; Bishop e Minasny, 2006; Grunwald,

2009). O objetivo principal da maioria dos trabalhos publicados até então é fornecer dados para a compreensão dos processos envolvidos na mitigação das emissões de carbono e aquecimento global (Grunwald, 2009). Segundo a autora é esse tipo de trabalho, centrado na qualidade ambiental, que atende às demandas sociais atuais. Assim, depois de período de desenvolvimento inicial, o MDS estaria se tornando cada vez mais um processo dirigido pelas necessidades dos usuários da informação (*user-driven*) (Carré et al., 2007; Finke, 2012).

2.1.3 O mapeamento digital de solos no Brasil

2.1.3.1 Características

O uso do MDS cresce no Brasil como cresce em outros países. Isso pode ser verificado, mesmo que de maneira grosseira, na Figura 2, que apresenta os resultados de uma busca que realizei no Google Acadêmico. Solicitei que fossem listadas todas as ocorrências de endereços na internet de páginas em português que contivessem a expressão "mapeamento digital de solos" (Pesquisa avançada do Google Acadêmico; Encontrar artigos com a frase exata "mapeamento digital de solos"; Incluir citações; Pesquisar somente nos idiomas selecionados → português). O mesmo foi feito para a expressão "digital soil mapping" para páginas em inglês. O período de busca foi definido como sendo o ano de 2003 por ter sido aquele em que as bases teóricas do MDS foram definidas. De lá para cá partimos de nenhuma (zero) ocorrência para 25 ocorrências, com a curva de tendência apresentando uma inclinação bastante significativa. Tal tendência é semelhante ao que ocorre no mundo todo, onde o número de ocorrências da expressão "digital soil mapping" é superior em apenas 10 vezes.

Contudo, o emprego do MDS no Brasil apresenta uma tendência divergente: a ênfase no mapeamento preditivo de classes de solo. De fato, a maioria dos trabalhos de predição espacial de solos realizados no Brasil até então envolveram a predição e mapeamento de propriedades do solo. Uma rápida pesquisa nos bancos de dados do Scielo e da ISI Web of Knowledge mostra essa realidade. Contudo, as predições têm se restringido a utilização da krigagem (e suas variantes) como função preditiva, onde os únicos fatores preditivos (co-variáveis) utilizados são o próprio solo (*s*) e as coordenadas espaciais (*n*). Geralmente esses trabalhos são desenvolvidos em áreas de pequena extensão (áreas experimentais, glebas de

produção) e, ao que tudo indica, não consideram que a abordagem utilizada é aquela do MDS (o modelo *scorpan* de McBratney et al., 2003). Apenas uma minoria utiliza expressões relacionadas ao MDS e/ou cita trabalhos considerados fundamentais para o seu desenvolvimento enquanto arcabouço teórico. Os exemplos mais recentes de trabalhos nessa área são aqueles de Mello et al. (2011) e Oliveira Junior (2011).

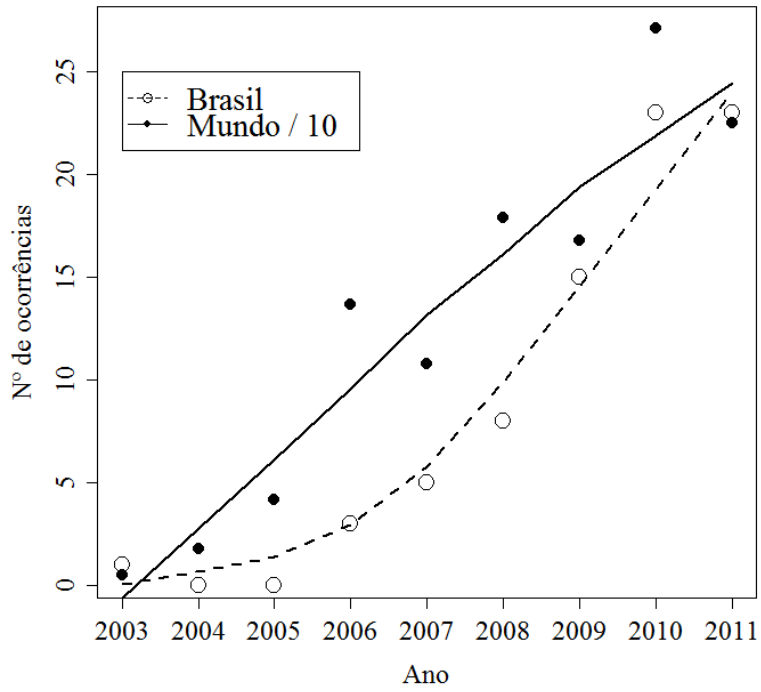


Figura 2 – Número de ocorrências das expressões “mapeamento digital de solos” e “digital soil mapping” no banco de dados do Google Acadêmico entre os anos de 2003 e 2011.

Até agora poucos trabalhos de mapeamento preditivo de propriedades do solo ousaram aplicar os fundamentos do MDS de maneira mais aprofundada no Brasil, valendo-se de modelos matemáticos mais complexos e/ou do uso combinado de modelos, o que permitiria a definição do componente determinístico da variação espacial das propriedades do solo (Bishop e McBratney, 2001). A maioria dos trabalhos ignora a possibilidade de valer-se de maior diversidade de variáveis preditoras (que são função dos fatores de formação do solo), de definir a incerteza associada às predições, ou mesmo aplicar os modelos a regiões mais extensas (bacias hidrográficas, regiões fisiográficas). Exemplos de aplicações mais “ousadas” são os trabalhos de Bernoux et al. (2006) e Mendonça-Santos et al. (2010), ambos empenhados em mapear os estoques de carbono no solo.

Como disse acima, a principal peculiaridade do MDS no Brasil é o seu uso crescente para a predição de classes de solo (ten Caten, 2011) desde 2006 quando foi publica o primeiro trabalho nacional com esse objetivo por Giasson et al. (2006). Depois desse trabalho outros nove foram publicados (ten Caten, 2011). Esses sim se consideram dentro do escopo do MDS, ao contrário da maioria dos trabalhos de predição espacial de propriedades do solo realizados no Brasil. Todos eles utilizam a expressão "mapeamento digital de solos" e/ou expressões correlatas no corpo do texto ou como palavras-chave. Além disso, sempre há referência a trabalhos que fundamentaram o MDS, indicando que os autores devem possuir bons conhecimentos sobre a área. Essa característica se deve, provavelmente, ao fato de que os pesquisadores que se envolveram na realização desses trabalhos têm, em sua maioria, formação de pedólogo ou estreita ligação à área da pedologia. Já no caso dos trabalhos de predição espacial de propriedades do solo, a maioria não tem a participação de pedólogos, mas, sobretudo de pesquisadores das áreas de física, fertilidade e manejo e conservação do solo. Isso mostra que a comunidade de pedólogos brasileiros começou a se interessar apenas tardiamente pela utilização de uma abordagem matemático-estatística (mais robusta) em suas atividades.

O trabalho de revisão de literatura realizado por ten Caten (2011) também mostra que no tocante a avaliação das FPESe desenvolvidas no Brasil, poucos trabalhos utilizam dados atualizados de campo. A grande maioria ainda se preocupa em avaliar a capacidade da FPESe reproduzir o mapa de solos elaborado através do método tradicional. Mas Grunwald (2009) mostra que, em termos mundiais, houve uma redução significativa do número de trabalhos preocupados em comparar, testar e/ou validar as FPESe utilizando mapas de classes de solo obtidas através do método tradicional, sobretudo por não ser uma abordagem de validação estatisticamente adequada (ten Caten, 2011). Mesmo assim, ela ainda é a mais utilizada no Brasil. Além disso, os pedólogos brasileiros têm construído FPESe para mapear solos nos quatro níveis categóricos do Sistema Brasileiro de Classificação de Solos (SiBCS). Para ten Caten (2011) a distinção e espacialização dessas classes na paisagem pode ser mero fruto do acaso, e não da capacidade preditiva das FPESe. Isso porque as propriedades do solo utilizadas para classificação nos níveis categóricos mais baixos do SiBCS como presença de mudança textural abrupta, concreções, espessura do horizonte A ou A+E, entre outras (Santos et al., 2006) seriam pouco relacionadas aos fatores de formação do solo (ten Caten, 2011) expressos através das variáveis preditoras usadas na construção das FPESe. Além disso, ten Caten (2011) observaram que tem crescido o uso de modelos mais complexos como as árvores de decisão e redes neurais artificiais na predição espacial das classes de solos. Apesar

da robustez e qualidade dos mapas preditos, os autores mostram-se preocupados com o fato dos estudos que utilizam essas técnicas têm dado pouca atenção às regras de decisão geradas durante a modelagem. Para eles não tem havido preocupação em esclarecer as relações existentes entre os fatores de formação do solo e as classes de solo preditas. Fenômeno semelhante foi observado por Grunwald (2009) em trabalhos publicados nos periódicos Geoderma e Soil Science Society of America Journal. Segundo a autora, os trabalhos que mostram elevada complexidade quantitativa geralmente apresentam uma significativa perda em interpretação pedológica e ambiental dos resultados, o que pode limitar a aplicação prática dos resultados. É por esse motivo que Minasny e McBratney (2007) afirmam que a melhoria da qualidade das predições de propriedades do solo depende muito mais da qualidade dos dados utilizados do que da complexidade dos modelos matemáticos.

2.1.3.2 Entendendo o mapeamento digital de solos no Brasil

O MDS foi adotado tardiamente pelos pedólogos brasileiros. Enquanto na Europa o mapeamento de classes de solo através do uso de funções de predição espacial já era realizado na década de 1970 (Webster e Burrough, 1972), no Brasil essa abordagem só foi aparecer na metade da década passada (Giasson et al., 2006). Para ten Caten (2011) essa adoção tardia está relacionada à disponibilização e popularização das ferramentas utilizadas no MDS (softwares, sistemas de informações geográficas, sensoriamento remoto) tardiamente no país, à carência de pessoal qualificado e ao conservadorismo dos pedólogos. Contudo, os motivos e as consequências dessa adoção tardia ainda não foram discutidas abertamente pela comunidade de pedólogos brasileiros. Por isso, realizei uma pesquisa com os pedólogos brasileiros que adotaram o MDS, bem como com os editores dos periódicos em que os trabalhos desses pedólogos foram publicados. Os questionários podem ser vistos no Anexo 1.

Retornemos aos motivos que determinaram a adoção tardia do MDS no Brasil. De fato, a disponibilização e popularização das ferramentas utilizadas no MDS, bem como dos conhecimentos necessários para tal, se deve a dificuldade de acesso às tecnologias e trabalhos desenvolvidos no exterior nas décadas passadas. A informática só se consolidou no meio acadêmico brasileiro a partir da década de 1990 com a maior abertura ao mercado externo. E somente em 2000, com a criação do Portal de Periódicos da Capes, boa parte dos trabalhos publicados no exterior se tornou disponível a comunidade acadêmica nacional. Mesmo assim,

nossa dificuldade em dominar a língua inglesa ainda nos impede de acessar boa parte desse conhecimento (sem falar no analfabetismo funcional que deve atingir parte significativa dos estudantes universitários (Vier, 2010)). Um relatório do índice de proficiência em inglês publicado por uma agência de intercâmbios mostra que, de uma lista de 44 países avaliados, o Brasil está na 31ª colocação, sendo classificado como de proficiência baixa (EF, 2011). Isso significa que muitos pesquisadores podem estar alheios a maior parte das publicações sobre MDS. Preocupados com esse cenário, alguns pedólogos brasileiros dão preferência a publicação de seus trabalhos na língua portuguesa. Segundo um dos pesquisadores que entrevistei, Cesar da Silva Chagas, isso é fundamental porque o MDS no Brasil apresenta características e necessidades distintas daquelas dos países em que tal abordagem está mais evoluída. Sabemos que essas diferenças são fruto do desenvolvimento desigual de nossa economia e da economia de países da América do Norte, Oceania e do Oeste Europeu. Léon Trotsky já mostrou que o fato de o capitalismo alcançar diferentes estágios de desenvolvimento em diferentes países determina o acesso aos meios de produção e a percepção do meio pela população. Por isso é que em países com maior disponibilidade de recursos e onde as necessidades básicas da população já estão satisfeitas, as preocupações com a qualidade ambiental são maiores (Kitamura, 1993). Assim, apesar de publicações em inglês terem uma visibilidade potencial maior, há que se ter o cuidado de não fazer ciência "para inglês ver" e assim contribuir para nos distanciarmos (ainda mais) da sociedade brasileira (Camargo e Inda Junior, 2011; Gatiboni, 2011).

As limitações com a língua inglesa e dificuldade de acesso ao conhecimento produzido no exterior podem estar relacionados com a carência de pessoal qualificado. O principal obstáculo é a falta de conhecimentos em matemática e estatística, sobretudo no que diz respeito às novas técnicas de MDS recém desenvolvidas por pesquisadores estrangeiros. Infelizmente não existem documentos publicados mostrando esse panorama. Contudo, com base em conversas informais, mensagens de e-mail trocadas e constatações feitas no exterior por Webster (2001) e Lark et al. (2007) de que muitos pedólogos ignoram matemática e estatística, é possível afirmar que essa é a realidade da pedologia brasileira. Isso significa que o corpo de profissionais habilitados a desenvolver estudos em MDS no Brasil é bastante limitado, assim como o de profissionais habilitados para avaliar projetos e artigos dessa área. Segundo alguns dos pedólogos que entrevistei, muitos editores e revisores científicos de periódicos nacionais desconhecem boa parte da terminologia estatística e matemática utilizada no MDS. Mas por que há carência de pessoal qualificado, principalmente em estatística e matemática, entre os pedólogos brasileiros? Provavelmente porque a maioria dos pedólogos

brasileiros tem formação agronômica, curso esse em que não há grande ênfase nas disciplinas de estatística e matemática. Trata-se do problema denunciado por Hartemink e McBratney (2008) e Grunwald (2009) de que estivemos sempre muito ligados estritamente a produção agrícola. Para Maria Leonor Lopes Assad a situação é ainda pior, pois muitos docentes da área de solos dos cursos de graduação apenas perpetuam "saberes cotidianos com roupagens de saberes científicos", o que seria agravado pela pós-graduação acelerada e produtivista que impossibilita uma sólida formação científica (Espindola, 2008).

Quanto à existência de conservadorismo por parte dos pedólogos brasileiros não há dúvidas. Um trabalho submetido por César da Silva Chagas a periódico nacional foi negado para publicação sob a alegação de que o MDS não constitui uma abordagem correta, por não atender às exigências do levantamento de solos tradicional, e que os resultados apresentados pelo trabalho representavam um demérito da pedologia tradicional. Segundo o pesquisador, dos três pareceres dados pelos revisores, apenas um foi negativo, mas mesmo assim o trabalho foi recusado com base nesse parecer contrário. Isso faz alguns pedólogos brasileiros apontarem o conservadorismo como o maior entrave ao desenvolvimento do MDS. Alguns apontam, inclusive, a necessidade de criação de espaços maiores nos periódicos nacionais, permitindo a consolidação do MDS no Brasil. Contudo, é preciso atentar que a atitude conservadora é perfeitamente compreensível, uma vez que nunca foi estranha a prática científica (Feyerabend, 1977). O que isso significa em termos epistemológicos? Significa que o modelo clorpt (Jenny, 1941), no qual o mapeamento tradicional de solos é baseado, deixou de ser uma teoria para se tornar uma doutrina. Uma teoria constitui um sistema de ideias aberto a toda informação que não é conforme a ela e que, portanto, pode questioná-la, enquanto que uma doutrina constitui um sistema de ideias fechado para toda informação não conforme (Morin, 1986). Como em toda situação, uma teoria dominante tende a se fechar em doutrina, as evidências que a suportam tornam-se verdadeiros dogmas, o que lhes permite contestar ao máximo a teoria nova que a contesta (Morin, 1986). E no caso da pedologia esse fenômeno parece ser ainda mais pronunciado. A passagem de Leeper (1956) reflete essa característica: "quando cientistas do solo discutem métodos para analisar [...] uma solução eles são práticos, racionais e não emotivos. Mas quando [...] discutem a classificação dos solos essas virtudes são suscetíveis de evaporar". A passagem mostra quanto os fatores irracionais costuma ser exacerbados entre nós pedólogos. E que pedólogo não gosta das palavras de Richard W. Arnold de que a pedologia é o coração, a alma e a habilidade artística da ciência do solo (Espindola, 2008)?

Esse mesmo conservadorismo pode ser o responsável pelo fato de os pedólogos

brasileiros que utilizam o MDS estarem preocupados em mapear classes de solo, e não propriedades do solo. Leeper (1965) diria que o "apego" ao SiBCS pode ser, para esses profissionais, uma questão de prestígio nacional. Além disso, o MDS brasileiro estaria passando por uma fase em que é preciso convencer os pedólogos conservadores de que a abordagem quantitativa é eficiente. Nada melhor do que usar os mapas de solos produzidos através do método tradicional como referência para validação das FPESe, mesmo que tal abordagem seja estatisticamente incorreta (ten Caten, 2011). O uso de tais comparações é fundamental para o progresso do MDS, uma vez que despertam elementos irracionais, muito mais eficazes do que a argumentação racional quando se pretende fazer aceitar uma nova teoria (Russell, 1932; Feyerabend, 1977).

Mas o foco no mapeamento de classes de solo também pode estar relacionado à maneira como o MDS foi introduzido no Brasil. Ramos (2003) afirmou no início da década passada que a pedologia estava entrando em um período de renovação, e que o MDS aparecia como um dos principais responsáveis. Tal notícia foi capaz de recuperar o ânimo de muitos pedólogos há muito tempo preocupados com a falta de financiamentos em face de uma realidade desagradável: apenas 35% do território nacional é coberto por mapas de solos em escala intermediária (1:100.000 a 1:600.000) (Mendonça-Santos e Santos, 2006). Apesar de Ramos (2003) destacar que a função do pedólogo vai muito além do simples mapeamento de solos, ele afirma que durante os anos por vir deveria permanecer no foco do pedólogo a organização e mapeamento de solos em classes homogêneas. Assim, o MDS foi inserido no Brasil como sendo a oportunidade dos pedólogos brasileiros reaverem o inacabado programa de mapeamento (de classes) de solos (Mendonça-Santos e Santos, 2006). E essa tarefa pode ser perfeitamente concluída dentro de uma década, desde que haja financiamento. Contudo, é fundamental atentar para a preocupação de Finke (2012) de que além de produzir informação (mapas de solos), os pedólogos atuais devem estar preocupados em produzir conhecimento, caso contrário corre-se o risco de sofrer a mesma perda de atenção e financiamento como ocorreu a partir das décadas de 1970 e 1980 (Finke, 2012). Grunwald (2009) e ten Caten (2011) já atentaram para a perda em interpretação pedológica e ambiental dos resultados do MDS à medida que aumenta a complexidade dos modelos utilizados. Assim, devemos ter em mente que não basta um modelo ser estatisticamente conveniente, mas, sobretudo que o mesmo faça algum sentido quando avaliado do ponto de vista pedológico (Milne e Lark, 2008).

Apesar dessas limitações e preocupações, diversas iniciativas vêm sendo tomadas para consolidar o MDS no Brasil. No meio acadêmico, devido à introdução tardia do MDS no

Brasil, ainda não existe nenhuma disciplina nos programas de pós-graduação que aborde esse tema. Quando muito, o MDS constitui um pequeno tópico dentro de outras disciplinas. Mas a Universidade Federal Rural do Rio de Janeiro acaba de realizar concurso para professor adjunto na área de concentração de Mapeamento Digital de Solos e Pedometria Aplicada (UFRRJ, 2011). Espera-se que com o concurso realizado na UFRRJ seja ofertada, já em 2012, a primeira disciplina do Brasil voltada especificamente para a pedometria e o MDS. Quanto aos grupos de pesquisa, o mais antigo registrado no CNPq data de 2000 (Pedometria e Mapeamento Digital de Solos), sendo sediado na Embrapa Solos. A principal atividade desenvolvida pelo grupo foi a organização do Second Global Workshop on Digital Soil Mapping, realizado no Rio de Janeiro em 2006. Logo em seguida foi criado o consórcio GlobalSoilMap.net. Como co-responsável pelo pólo da América Latina e Caribe do GlobalSoilMap.net, juntamente ao Centro Internacional de Agricultura Tropical (CIAT), a Embrapa Solos assumiu o grande compromisso junto ao consórcio GlobalSoilMap.net de mapear os solos da América Latina e Caribe. Mais recentemente a Embrapa Solos encabeçou a criação da RedeMDS. Financiada pelo CNPq e EMBRAPA, a RedeMDS conta com pesquisadores de diversas instituições do país e tem por objetivo elaborar projetos para o MDS do Brasil em alta resolução, inserindo-se no projeto mundial GlobalSoilMap.net (Dias, 2011).

Mas infelizmente o suporte financeiro para os projetos de MDS no Brasil e atividades da RedeMDS não parece ser tão significativo (e suficiente) como ocorre em outros países. Atualmente, os maiores investimentos são realizados na América do Norte, Europa, Austrália e África, sendo que na última os recursos são predominantemente estrangeiros. A preocupação em financiar os trabalhos de MDS da África tem sido atribuída ao fato de que o conhecimento a respeito dos solos daquele continente é muito restrito. Segundo Pedro Sanchez, “nós sabemos mais sobre os solos de Marte do que sobre os solos da África”. Nesse sentido, em novembro de 2008, 18 milhões de dólares foram obtidos da fundação Bill & Melinda Gates e da Aliança para uma Revolução Verde na África (AGRA) como financiamento para mapear a maior parte da África sub-saariana (GlobalSoilMap.net, 2011).

Mas por que motivo não há suporte financeiro para mapeamentos sistemáticos de solos no Brasil, diferente do que ocorreu nas décadas de 1970 e 1980? Para entender esse cenário é preciso fazer uma recapitulação histórica. Devemos lembrar que as décadas de 1970 e 1980 foram marcadas por Planos de Desenvolvimento Nacional (PDNs) embasados nas ideias de crescimento acelerado do tecnocrata Delfim Neto (Macarini, 2005) (Delfim Neto que era grande conhecedor dos solos da Amazônia: “existe na Amazônia uma mancha de terra roxa

comparável à de qualquer estado da região Centro-Sul" (Skidmore, 1989)). Um dos focos dos PDNs era o incentivo a agricultura através da abertura de novas áreas, ampliação da mecanização e do uso de insumos modernos. Projetos como o Radam e o RadamBrasil, e os levantamentos de solo realizados nas duas décadas, permitiram a colonização das regiões Norte, Nordeste e Centro-Oeste (Silva, 2005). Foi a consolidação da Revolução Verde. Mas como uma reforma agrária não foi implantada concomitantemente, a Revolução Verde foi apenas um instrumento de "modernização conservadora" que ajudou a aprofundar e agravar as desigualdades no campo (Brum, 1988). Para alimentar o "milagre econômico" foram necessários recursos dos quais o país não dispunha, levando ao aumento da dívida pública e consequente necessidade de empréstimos junto ao Fundo Monetário Internacional (FMI). Em 1982 a situação da economia brasileira se tornou insustentável, em 1985 o FMI suspendeu a ajuda financeira ao Brasil e em 1993 a inflação chegou a 2.700% (Faria, 2007). É a crise financeira apontada por Ramos (2003) como responsável pela redução no incentivo aos levantamentos de solos. Mas àquele tempo os levantamentos de solos já haviam cumprido com a sua missão, que era a de promover políticas governamentais regionalizadas (Espindola, 2008), sobretudo a ocupação agropecuária das regiões Centro-Oeste, Norte e Nordeste.

O panorama atual é bem diferente do período da ditadura militar. O principal foco do governo é a eliminação da pobreza: "País rico é país sem pobreza". Nada mais adequado em um país que possui 16,27 milhões de pessoas em extrema pobreza, das quais 46,7% residem no meio rural (IBGE, 2010), fruto do modelo de desenvolvimento adotado nas décadas passadas. Assim, se os grandes projetos de levantamentos de solos (e outros recursos naturais) das décadas de 1970 e 1980 só foram possíveis graças a um modelo de desenvolvimento que favoreceu os interesses econômicos de grandes corporações e agravou as desigualdades sociais. Se o "sucesso" desse modelo se deve, em grande parte, aos levantamentos de solos realizados. Como a realização de novos mapeamentos de solos pode ajudar, hoje, a eliminar a pobreza? Já foi demonstrado que a segurança alimentar será uma realidade somente através de políticas públicas de redistribuição de renda, garantindo o acesso aos meios de produção e serviços básicos (Silveira e Almeida, 1992; Monteiro, 2003). A FAO já reconheceu que "o problema não é tanto a falta de alimentos, mas a falta de vontade política" (FAO, 2005). E se os hábitos de consumo e aumento da população são os maiores responsáveis pela pegada ecológica humana (Dietz et al., 2007), então somente a mudança dos hábitos de consumo e o controle populacional poderão ter efeito significativo em questões como a mudança climática, a degradação ambiental, a escassez de água e as ameaças a biodiversidade. Mesmo assim continuamos justificando a realização de novos projetos de mapeamento de solos (bem como

outros projetos de pesquisa sobre solos) com o argumento de que são *fundamentais* para resolver os principais problemas sociais e ambientais do nosso tempo. Se nossas atividades científicas estivessem, como propõe Feyerabend (1977), submetidas ao controle da sociedade, do indivíduo leigo, seriam elas consideradas meritórias de receber financiamentos? Precisamos refletir sobre essa questão! Eu, particularmente, creio que não existe uma demanda significativa no Brasil, hoje, para informações sobre solos, mesmo que nós pedólogos consideremos necessário. E sem demanda não haverá financiamento. Lembremos que o trabalho seminal de Vasily Dokuchaev só foi possível porque o governo se preocupou com uma seca que ocorreu em uma das maiores regiões produtoras de cereais da Rússia (Espindola, 2008). Além do mais, devido ao crescimento da economia brasileira ocorrido nos últimos anos, a comunidade internacional espera que não sejam necessários recursos externos para financiar o desenvolvimento dos nossos projetos de mapeamento (digital) de solos.

2.1.3.3 O que podemos fazer?

Os pedólogos brasileiros estão, gradativamente, aderindo ao MDS. Isso significa que cresce o interesse da comunidade científica nacional em utilizar tal abordagem na produção de informações e conhecimentos sobre o solo. Mas para isso precisamos formar pessoal qualificado. Essa qualificação deve contemplar conhecimentos matemáticos e estatísticos, mas, principalmente, pedológicos. E é por esse motivo que os pedólogos conservadores também são fundamentais. Eles devem estar atentos para que os produtos do MDS não se tornem apenas produtos, informação, mas que gerem e aperfeiçoem nosso conhecimento a respeito do solo e suas interações com o ambiente. Precisamos de modelos matemáticos com significado pedológico. Mas para que possamos atingir esse nível é fundamental a reformulação dos currículos dos cursos de graduação que formam os futuros pedólogos (agronomia, engenharia florestal, engenharia agrícola, geografia, entre outros), melhorando a formação pedológica e dando mais ênfase a estatística e matemática. O mesmo deve ocorrer com os programas de pós-graduação em ciência do solo. Estes ainda deverão dar mais abertura aos profissionais de outras áreas, como as engenharias, geografia, matemática, estatística, entre outras, deixando de olhar apenas para dentro das ciências agrárias (sobretudo a agronomia). Essa qualificação ainda passa pela necessidade dos estudantes de graduação e pós-graduação dominarem a língua inglesa, não para publicar nossos artigos em inglês, mas

sim para acessar o conhecimento produzido no exterior.

Também precisamos identificar quem são os verdadeiros usuários potenciais das informações e conhecimentos que produzimos através do MDS. Somente dessa maneira será possível evoluir em direção a uma pesquisa mais *user-driven*. Até o presente momento a maioria dos projetos desenvolvidos no Brasil parece ter sido elaborada para responder às curiosidades e preocupações dos pedólogos (como é o caso do projeto que originou a presente dissertação de mestrado). Somente me preocupo com isso porque utilizamos recursos públicos. Tendo definido os potenciais beneficiados pelo MDS poderemos definir que tipo de informação e conhecimento deve ser produzido (devemos partir do mapeamento de classes ou propriedades do solo?). E, mais do que isso, avaliar quais serão as consequências de nossas pesquisas, tentando eliminar o "componente ilusório de nossa percepção" (Morin, 1986) de que os produtos do MDS são fundamentais para solucionar os principais problemas sociais e ambientais de nosso tempo.

Além disso, não podemos nos esquecer de tentar, dentro do possível, padronizar nossos métodos e procedimentos de mapeamento de solos. A diversidade de métodos e procedimentos já era comum no procedimento mapeamento tradicional de solos e parece estar presente também no MDS. Historicamente isso causou problemas na organização e comparação dos dados. O documento contendo as especificações que definem os aspectos a serem obedecidos para permitir o agrupamento e a apresentação de produtos finais padronizados do consórcio GlobalSoilMap.net constitui um bom ponto de partida. Tais especificações definem três aspectos (GlobalSoilMap.net, 2011):

a) As (duas) entidades espaciais de trabalho: a entidade espacial primária constitui uma célula volumétrica (volumetric pixel – voxel) com dimensões horizontais de 100 m por 100 m localizada no ponto central de um grid global de 3 arcos-segundo por 3 arcos-segundo (aproximadamente 93 m no equador). Na dimensão vertical, as predições dos valores das propriedades do solo e suas incertezas associadas serão feitas até 2 m com dados reportados para seis intervalos de profundidade (0-5 cm, 5-15 cm, 15-30 cm, 30-60 cm, 60-100 cm e 100-200 cm). A entidade espacial subsidiária constitui um ponto com dimensões horizontais irregulares de não mais de 2 m por 2 m localizado no centro do mesmo grid global de 3 arcos-segundo por 3 arcos-segundo. Os valores preditos para as entidades espaciais subsidiárias são agregados para produzir as predições de bloco (entidade espacial primária). A localização deverá ser feita utilizando a projeção geográfica e o WGS84 como datum horizontal (localização espacial), bem como o ano (localização temporal).

b) As doze propriedades do solo prioritárias: (a) profundidade total do perfil (cm), (2) profundidade efetiva do solo (cm), (3) carbono orgânico (g kg^{-1}), (4) pH (x10), (5) areia (g kg^{-1}), (6) silte (g kg^{-1}), (7) argila (g kg^{-1}), (8) cascalho (% vol), (9) capacidade de troca de cátions efetiva ($cmol_c$ kg^{-1}), (10) densidade da fração terra fina (< 2 mm) (sem cascalho) (Mg m^{-3}), (11) densidade do solo inteiro *in situ* (inclui cascalho) e (12) capacidade de água disponível (mm). Essas propriedades devem ser determinadas através de métodos específicos (padrão), sendo necessário desenvolver modelos para ajustar os resultados de outros métodos analíticos.

c) A avaliação das FPESe: Cada propriedade do solo deve ter uma estimativa da incerteza associada com a predição para cada profundidade para cada local no grid. A incerteza é definida como os 95 % do intervalo de predição. A medida apropriada de acurácia para cada propriedade do solo é o quadrado médio do erro para as predições pontuais (2 m x 2 m). Essa estatística pode ser obtida através do uso da validação-cruzada para métodos baseados em pontos e validação de campo para métodos baseados em classes de solo.

Cabe avaliarmos se as especificações do GlobalSoilMap.net (2011) podem ser utilizadas como modelo em busca da padronização do MDS no Brasil. Essa avaliação é fundamental, principalmente se considerarmos que o MDS no Brasil possui características e necessidades distintas daquelas dos países em que tal abordagem está mais evoluída (Cesar da Silva Chagas, comunicação pessoal). Além disso, devemos estar atentos ao fato de que tais especificações são inconsistentes do ponto de vista pedogenético. A definição de intervalos de profundidade de coleta "arbitrários" claramente desconsidera a morfologia do solo e sua variabilidade vertical. Trata-se do mesmo "erro" apontado no método tradicional de mapeamento de solos que desconsidera a variabilidade horizontal do solo. Mas note que a definição dos seis intervalos de profundidade de coleta em detrimento do uso dos horizontes do solo se deve, entre outros aspectos, a recomendação de que os mapas de solos devem dar menos ênfase à taxonomia. Além disso, ainda não existem métodos que permitam a interpolação e visualização adequada de dados de solos em três dimensões.

O relatório do GlobalSoilMap.net (2011) ainda apresenta outras especificações, mas uma das mais interessantes é o aparente incentivo ao uso de softwares livres como o R (R Development Core Team, 2011). Alguns algoritmos para uso nos projetos do consórcio estão sendo desenvolvidos para serem implementados no R. A maior vantagem do R em relação aos softwares comerciais, além de ser gratuito, é a possibilidade de que as rotinas de análise sejam registradas e, quando necessário, verificadas ou disponibilizadas para verificação por outros

pesquisadores. Tal possibilidade é fundamental para uma prática científica que se julgue séria. Isso reforça a necessidade de substituição de pacotes estatísticos pagos por aqueles de código aberto nas universidades públicas brasileiras. Sobretudo pelo fato de que não há necessidade de gastarmos recursos públicos com aquisição de softwares se temos produtos gratuitos com capacidade de trabalho equivalente e até superior.

Por fim, precisamos determinar quais são os custos do MDS, conforme já foi apontado por Grunwald (2009) e ten Caten (2011). E, mais do que isso, quantificar os benefícios (sociais, econômicos, culturais, ambientais) que os produtos do MDS podem gerar. Que tipo de produto traz os maiores benefícios: mapas de classes ou mapas de propriedades do solo? Devemos fornecer à sociedade critérios sólidos e racionais que justifiquem o financiamento dos projetos de MDS no Brasil, deixando de lado as especulações emocionais de que o MDS é fundamental para resolver os principais problemas sociais e ambientais de nosso tempo.

2.2 Funções de predição espacial de propriedades do solo

O MDS é realizado através do uso de funções de predição espacial de solos com erros autocorrelacionados espacialmente (FPESe). Tais funções constituem um método de ajuste empírico das relações quantitativas existentes entre o solo e o ambiente em que o mesmo ocorre (McBratney et al., 2003). A estrutura das FPESe é do tipo

$$S = f(s, c, o, r, p, a, n) \quad (1)$$

que constitui o modelo *scorpan*, onde S = classe ou propriedade do solo a ser predita em função de s = informação do solo previamente disponível, c = clima, o = organismos, r = relevo, p = material de origem, a = tempo, e n = posição espacial (McBratney et al., 2003). Esse modelo constitui uma generalização do bastante conhecido modelo *clorpt* (Jenny, 1941):

$$S = f(cl, o, r, p, t, \ldots) \quad (2)$$

onde *S* = solo, função de *cl* = clima, *o* = organismos, *r* = relevo, *p* = material de origem, e *t* = tempo, além de fatores desconhecidos (...). A principal diferença na estrutura das duas formulações é a existência, no modelo *scorpan*, dos fatores *n* e *s*. O fator *n* espacializa as relações entre o solo e o ambiente em que o mesmo ocorre, indicando que o solo pode ser predito a partir de coordenadas geográficas. Além disso, o próprio fator adicional *s*, que indica que o solo pode ser predito a partir de si mesmo, possui, implicitamente, coordenadas espaciais (McBratney et al., 2003). Essa abordagem espacial permite predizer propriedades do solo levando em consideração um dos aspectos mais importantes para o seu funcionamento como componente ambiental: a variabilidade espacial. Além disso, os produtos do MDS incluem os conceitos de incerteza e acurácia associados às FPESe (McBratney et al., 2000), o que não é encontrado no modelo *clorpt*.

Um aspecto importante a ser esclarecido aqui é a comum confusão feita entre os conceitos de FPESe e de função de pedotransferência. De fato, os dois conceitos são correlacionados, havendo sobreposições, mas diferenças podem ser facilmente demonstradas. Segundo McBratney et al. (2003), a principal delas é que a estrutura de uma função de pedotransferência é do tipo *s=f(s)*, ou seja, uma propriedade do solo é sempre função de outra(s) propriedade(s) do mesmo solo, sem que haja, necessariamente, referência espacial. Assim, uma função do tipo *s=f(r)* não constitui uma função de pedotransferência, mas sim uma FPESe.

Mas para que uma FPESe possa ser construída dependemos, em primeiro lugar, da existência de um conjunto de dados a partir do qual estabelecemos as relações entre o solo e as variáveis ambientais utilizadas como preditoras. Para isso geralmente é necessário realizar a amostragem dos solos que, em termos gerais, pode ser conduzida de duas maneiras. Na primeira delas amostramos a área sob investigação em toda a sua extensão, sendo os pontos amostrais definidos a partir de algum critério a priori (Webster e Oliver, 1990; McBratney et al., 2003). Já na segunda, amostramos apenas uma pequena área que seja representativa de toda a área sob investigação, a qual constitui a área de referência (ten Caten et al., 2011c).

Vários são os critérios que podemos utilizar para definir a maneira como as amostras são coletadas na área sob investigação quando amostrada em toda a sua extensão. Uma delas, e talvez a mais recomendada, mas nem por isso a mais praticada, é a amostragem aleatória (Brus e Gruijter, 1997). Através desse procedimento todos os pontos da área (pixeis de um mapa digital) têm a mesma probabilidade de serem amostrados, sendo por isso chamada de amostragem probabilística (Brus et al., 2011). Outra possibilidade é a amostragem dos solos utilizando uma malha regular de pontos amostrais (Lark e Bishop, 2007), o que talvez seja o

procedimento mais comum, sobretudo em estudos diretamente relacionados à agricultura de precisão. Ambos os casos podem ser realizados após a estratificação da área sob investigação com base em variáveis categóricas como geologia, uso da terra, classe de solo, unidades geomorfológicas, entre outras. O número de amostras pode ser o mesmo em cada categoria (Gessler et al., 1995) ou proporcional à sua área de abrangência (Heim et al., 2009). Variações e adaptações desses procedimentos amostrais podem ser encontradas em McKenzie et al. (2008).

Através dessa primeira forma de amostragem, onde toda a área sob investigação é amostrada, o objetivo é coletar toda a variação existente na área, ou seja, abranger todo o intervalo de valores de cada variável de interesse (McBratney et al., 2003). Isso possibilita que a FPESe construída trabalhe sem ter que extrapolar para além dos seus limites (McBratney et al., 2003). Seu desempenho seria, nesse caso, maximizado. Em geral, essa abordagem amostral se aplica aos casos em que a área sob investigação não apresenta problemas relacionados ao acesso por não haverem impedimentos fisiográficos e/ou financeiros. Além disso, em áreas de tamanho reduzido é mais fácil fazer uma amostragem intensiva, como o fez Samuel-Rosa et al. (2011a) (uma amostra a cada 0,07 ha). As FPESe são então construídas através do uso de modelos estatísticos (modelos lineares, modelos lineares generalizados, árvores de regressão e decisão, entre outros), geoestatísticos (como krigagem e co-krigagem) ou uma combinação dos dois (métodos híbridos) (Webster e Oliver, 1990; McBratney et al., 2003; Bishop e Minasny, 2006; Grunwald, 2009).

Contudo, em muitas situações, talvez na maioria delas, possuímos restrições de ordem financeira e de força de trabalho, o que impede a realização de amostragens sistemáticas em toda a extensão da área sob investigação. Esse é um dos principais motivos pelos quais grandes extensões do planeta ainda não possuem seus solos mapeados. Para os casos em que o número de amostras capazes de serem obtidas é limitado, Brus e Heuvelink (2007) tentaram identificar um padrão amostral ótimo que possibilitasse minimizar o erro das estimativas. Contudo, a área de investigação também pode ser de difícil acesso, tornando inviável a amostragem em toda a sua extensão. Como forma de contornar esses problemas podemos lançar mão da segunda abordagem amostral a que me referi acima, que se utiliza do conceito de área de referência. Esse conceito foi desenvolvido no início da década de 1980 na França, mas veio a ser associado ao modelo *scorpan* (que ainda não havia sido formalizado) somente na década seguinte por Lagacherie et al. (1995). Segundo o Lagacherie et al. (1995) existem regiões naturais que possuem características similares em termos de geologia e topografia, o que condiciona a ocorrência de solos em um padrão também similar e que se repete na

paisagem. O conceito de região, província ou unidade fisiográfica utilizado pelo IBGE (2004) ("região caracterizada por elementos da estrutura e natureza das rochas, acrescidos das indicações da rede hidrográfica, do clima, do aspecto topográfico e da idade das rochas") pode ser utilizado com alguma equivalência. Assim, delimitamos a menor região natural que possua as características e padrões que definem toda a região sob investigação, utilizando para isso informações dos fatores de formação do solo (Lagacherie et al., 1995). Nessa pequena região natural faz-se uma amostragem intensa do solo e de outras variáveis ambientais, a partir do que construímos as FPESe (Voltz et al., 1997). Contudo, é fundamental identificar a representatividade das áreas de referência (Lagacherie et al., 2001), ou seja, delimitar a região natural dentro da qual a FPESe é realmente capaz de capturar as relações existentes entre o solo e as demais variáveis ambientais. Lagacherie et al. (2001) conseguiram, com sucesso, delimitar essas regiões naturais no Sul da França utilizando distâncias matemáticas. De posse dos dados, mais uma vez as FPESe são construídas através do uso de modelos estatísticos (modelos lineares, modelos lineares generalizados, árvores de regressão e decisão, entre outros), geoestatísticos (como krigagem e co-krigagem) ou uma combinação dos dois (métodos híbridos) (Webster e Oliver, 1990; McBratney et al., 2003; Bishop e Minasny, 2006; Grunwald, 2009). De posse das FPESe faz-se a predição dos solos no resto da área sob investigação que não teve os solos amostrados, ou seja, onde dispomos apenas das variáveis preditoras (ten Caten et al, 2011c).

Na construção das FPESe os atributos de terreno são os mais utilizados como variáveis preditoras (McBratney et al., 2003; Bishop e Minasny, 2006; Grunwald, 2009; ten Caten, 2011), prática que encontra fundamentação no conceito de catena desenvolvido por Milne em 1936 (Grunwald, 2006). Esse conceito está relacionado à hipótese levantada por Moore et al. (1993): se a água é o agente erosional dominante em uma bacia hidrográfica e, portanto, possui uma importante função no desenvolvimento de topossequências de solos, então a distribuição espacial dos atributos de terreno que caracterizam os fluxos de água também captura a variação espacial das propriedades do solo. Além dessa estreita relação entre os atributos do terreno e o padrão de distribuição espacial dos solos, os atributos de terreno estão entre as variáveis ambientais de mais fácil obtenção (ten Caten, 2011). Eles são derivados de modelos digitais de elevação (MDEs) obtidos a partir de sensores remotos, pela digitalização de curvas de nível de cartas planialtimétricas ou através de levantamentos topográficos. Tais fontes de dados estão, geralmente, disponíveis em resolução espacial média a alta, enquanto que informações sobre material de origem e clima estão disponíveis em baixa resolução ou estão desatualizadas (tem Caten, 2011). Além disso, informações sobre o material de origem

costumam ser produzidas com base nas informações topográficas, como em Maciel Filho (1990).

A partir dos MDEs podemos derivar diversos atributos primários e secundários, havendo na literatura uma variação de mais de uma centena (Hengl e MacMillan, 2009). Os atributos primários são aqueles calculados diretamente a partir do MDE, como a elevação, a declividade, as curvaturas (de perfil e planar), o aspecto ou orientação, a área de contribuição, a área de contribuição específica, entre outros (Moore et al., 1993). Enquanto isso os atributos secundários são aqueles calculados a partir de dois ou mais atributos primários, como é o caso do índice de umidade topográfica, o índice de capacidade de transporte de sedimento, o fator LS, entre outros (Moore et al., 1993). Mais exemplos de atributos de terreno e o método de obtenção de cada um deles são apresentados abaixo na seção Material e Métodos. Informações ainda mais detalhadas podem ser encontradas no trabalho de Wilson e Gallant (2000b).

Em geral, dentre todos os atributos de terreno, os primários são aqueles mais utilizados na construção das FPESe, sobretudo a elevação e a declividade (McBratney et al., 2003; Bishop e Minasny, 2006). Talvez isso possa ser explicado pelo fato de que os atributos primários e secundários possuem uma relação intrínseca bastante forte, fazendo com que apresentem padrões similares, o que resulta na sobreposição de informações (Hengl e MacMillan, 2009). Obviamente isso ocorre porque os atributos secundários são construídos a partir dos atributos primários. Na construção de uma FPESe a existência de variáveis preditoras que "compartilham" parte significativa da variância predita constitui o que se chama de multicolinearidade, o que geralmente possui efeito negativo sobre a análise (Hair et al., 2010). Por esse motivo há a preocupação em utilizar métodos de análise multivariada de dados como a análise de componentes principais para identificar variáveis fortemente correlacionadas antes de construir as FPESe (ten Caten et al., 2011b). Nesse caso, os atributos primários parecem ser preferidos em detrimento dos atributos secundários devido ao fato de serem de mais fácil obtenção e compreensão.

Apesar da aparente facilidade de construção de FPESe, não é possível afirmar que possamos utilizá-las para mapear as propriedades dos solos de toda e qualquer superfície geomorfológica. O desempenho das FPESe é bastante variável e dificilmente ultrapassa coeficientes de determinação superiores a 70-75%. Moore et al. (1993) utilizaram modelos de regressão linear para construir FPESe para estimar o teor de silte e argila do solo e conseguiram explicar no máximo 66% da variância. Com resultados similares, Gessler et al. (1995), conseguiram explicar no máximo 68% da variância da profundidade do solum usando

modelos lineares generalizados. McKenzie e Ryan (1999) conseguiram explicar 42%, 78% e 54% da variância da profundidade do solo e dos teores de fósforo total e de carbono total, respectivamente, usando árvores de regressão e modelos lineares generalizados. Da mesma maneira, Gobin et al. (2001) tentaram predizer cinco propriedades de solos nigerianos usando modelos lineares múltiplos, onde conseguiram explicar entre 41% e 75% da variância. E Sumfleth e Duttmann (2008) não conseguiram explicar mais de 41% da variância da proporção de silte e do teor de carbono total em solos chineses. De qualquer forma, os desempenhos alcançados pelas FPESe construídas nesses trabalhos é similar, quando não superior, ao desempenho de mapas de solos produzidos através do método tradicional, que não passa de aproximadamente 50% (Webster e Oliver, 1990).

Essa incapacidade das FPESe em explicar mais do que 70-75% da variância se deve ao fato que a relação entre as propriedades do solo e as demais variáveis ambientais nem sempre é evidente: ela pode ser forte em uma superfície geomórfica, mas fraca em outra (Grunwald, 2006). Obviamente isso depende das variáveis preditoras utilizadas na construção das FPESe. No caso dos atributos de terreno, que são os mais utilizados, essa incerteza deriva do fato que algumas superfícies são menos estáveis ou sofreram alterações naturais e/ou antropogênicas recentemente (sem contar os erros associados ao MDE de onde os atributos de terreno são derivados). Isso as torna mais complexas, o que está diretamente relacionado à sua idade, podendo ser consideradas superfícies mais jovens (FAO, 2006). Como consequência direta, nessas superfícies geomórficas os fatores de formação do solo também acabam atuando de maneira complexa. Por outro lado, pode-se esperar que a relação entre as propriedades do solo e os atributos de terreno em superfícies mais velhas e estáveis seja máxima. Mas mesmo assim parece que não temos garantias. Os outros fatores de formação do solo, principalmente o fator humano, podem ter causado distúrbios suficientes (alterações antropogênicas que "rejuvenescem" os solos e as superfícies geomórficas → teoria da pedogênese reversa (Streck et al., 2008; Samuel-Rosa et al., 2011b) para fazer com que a relação seja mínima ou até mesmo inexistente.

Se diversos fatores afetam de maneira complexa a relação entre as propriedades solo e as demais variáveis ambientais, tornando impossível construir uma FPESe que possa ser utilizada em todas as superfícies geomorfológicas (Grunwald, 2009), é necessário desenvolvermos estudos nas mais diversas condições (McKenzie et al., 2000). Somente assim será possível (1) verificar se existe alguma relação entre as propriedades do solo e as variáveis ambientais para que possam ser construídas FPESe e (2) estabelecer os domínios fisiográficos dentro dos quais as FPESe construídas podem ser utilizadas. Uma dessas superfícies

geomórficas é o Rebordo do Planalto do Estado do Rio Grande do Sul (RS). A região constitui a zona de transição entre duas importantes regiões fisiográficas do RS: o Planalto e a Depressão Central. Sua geologia é complexa, incluindo diversas rochas ígneas extrusivas e sedimentares (Sartori, 2009), o que define a topografia e a hidrologia regionais. O relevo varia entre plano e montanhoso e deslizamentos de terra ocorrem periodicamente (Pinheiro & Soares, 2004). Os solos são frágeis, predominantemente arenosos e siltosos, configurando elevada suscetibilidade a erosão (Dalmolin & Pedron, 2009). Do ponto de vista histórico, essa região foi intensamente utilizada para agricultura, sobretudo a partir do final do século XIX, quando boa parte da vegetação nativa foi derrubada (Neumann, 2003). O uso indiscriminado de áreas impróprias para agricultura, associado a práticas não-conservacionistas de preparo do solo, resultaram na degradação de muitas áreas (Samuel-Rosa et al., 2011b). Esses aspectos evidenciam que o Rebordo do Planalto do RS possui constituição geológica e pedológica complexa e, devido às alterações naturais e antropogênicas recentes, constitui uma superfície geomórfica jovem e instável. Provavelmente os fatores de formação do solo não atuam de maneira uniforme, imprimindo dificuldades a construção de FPESe para estimar propriedades do solo. Não deve se esperar que em situações como essa, FPESe apresentem desempenho superior a 50 %, o que seria equivalente ao desempenho dos mapas tradicionais de solo.

Entretanto, nos casos em que a área sob investigação apresenta elevada complexidade geológica e pedológica, a recomendação é de que seja realizada a estratificação da região em domínios mais homogêneos antes de construir FPESe. Gessler et al. (1995) recomendam que essa estratificação seja realizada com base na geologia, uma vez que esse fator de formação do solo costuma possuir efeito significativo sobre as propriedades do solo. No caso do Rebordo do Planalto do RS, a estratificação em domínios fisiográficos mais homogêneos constitui uma possibilidade bastante plausível. Ela pode ser baseada no tipo de material de origem, vindo a constituir dois domínios fisiográficos: o primeiro onde predominam rochas ígneas extrusivas e o segundo onde predominam rochas sedimentares. Como solos derivados de rochas sedimentares possuem distribuição do tamanho de partículas mais grosseira do que os solos derivados de rochas ígneas extrusivas (Miguel, 2009), essa estratificação deve ter efeito positivo sobre o desempenho de FPESe. Talvez seja possível alcançar o desempenho máximo sugerido por Moore et al. (1993) que é de 70-75%.

2.2.1 Predição da distribuição do tamanho de partículas do solo

O principal uso de FPESe é para a predição de propriedades do solo (McBratney et al., 2003; Bishop e Minasny, 2006; Grunwald, 2009). Dentre todas as propriedades do solo, a mais importante é a distribuição do tamanho de partículas. Isso porque ela determina e influencia a maioria das demais propriedades do solo, como o comportamento hidráulico, a capacidade de armazenamento de água, a consistência, a formação e manutenção da estrutura, o ciclo dos diversos compostos químicos e íons, a suscetibilidade a erosão, entre outras (Reichert et al., 2003; Hillel, 2005). Como consequência, diversos são os trabalhos realizados onde uma das principais preocupações foi a predição da distribuição do tamanho de partículas do solo, como aqueles de Moore et al. (1993), Odeh et al. (1993), McBratney et al. (2000), Park e Vlek (2002), Sumfleth e Duttmann (2008), entre muitos outros.

A grande maioria das FPESe são construídas para predizer cada fração de tamanho de partícula (areia, silte, argila) separadamente. Entretanto, segundo Aitchison (1982), esse procedimento é incorreto, uma vez que a distribuição do tamanho de partículas do solo constitui o que se chama de dados composicionais. Dados composicionais são aqueles que apresentam informações relativas, ou seja, representam as partes de um todo. Nesse caso, a característica fundamental desse tipo de dados é que a sua soma sempre resulta uma constante, 1 para o caso de proporções e 100 para o caso de percentagens. Isso significa que esses dados são impedidos de variar no intervalo entre $-\infty$ e $+\infty$ conforme requerido pelos métodos estatísticos clássicos (Aitchison, 1982).

Contudo, poucos pesquisadores têm se preocupado com a predição da distribuição do tamanho de partículas do solo utilizando a abordagem de dados composicionais. Os exemplos mais importantes são os trabalhos de Odeh et al. (2003), Lark e Bishop (2007) e Rawlins et al. (2009). Segundo Pawlowsky-Glahn e Egozcue (2006) a maioria dos pesquisadores reluta em utilizar a abordagem de dados composicionais porque a análise das frações de tamanho separadamente e/ou utilizando os métodos estatísticos clássicos ainda fornecem resultados que permitem uma interpretação aparentemente coerente. É por esse motivo que não existe consenso entre os pesquisadores sobre a necessidade de análise dos dados da distribuição do tamanho de partículas como dados composicionais.

2.2.1.1 Análise de dados composicionais

A discussão sobre a análise de dados composicionais teve início em 1897 com o artigo de Karl Pearson sobre correlações espúrias (Buccianti et al., 2006). Mas a solução para a análise de tais dados apareceu somente no início da década de 1980 (Aitchison, 1982), culminando com a publicação do livro *The Statistical Analysis of Compositional Data* no ano de 1986 por John Aitchison. De lá para cá diversos pesquisadores têm trabalhado para desenvolver e popularizar essa teoria, dentre os quais se destacam o próprio John Aitchison, através de cursos de curta duração (Aitchison, 2003), e Vera Pawlowsky-Glahn, da Universidade de Girona, na Espanha. Essa autora participou da editoração da última grande contribuição coletiva para a teoria da análise de dados composicionais (Pawlowsky-Glahn e Egozcue, 2006), que constitui no 264º volume da Geological Society of London intitulado *Compositional Data Analysis in the Geosciences: From Theory to Practice* (Buccianti et al., 2006). Através dessas contribuições, descrevo a seguir as propriedades dos dados composicionais, as consequências da sua análise através dos métodos estatísticos clássicos e as maneiras mais adequadas para sua análise. Todas as informações apresentadas foram obtidas de Aitchison (1982), Aitchison (2003) e Pawlowsky-Glahn e Egozcue (2006).

2.2.1.1.1 Propriedades das composições e suas consequências

Dados composicionais são aqueles que apresentam informações relativas. A característica fundamental desse tipo de dados é que a sua soma sempre resulta uma constante (1 para o caso de proporções e 100 para o caso de percentagens). Isso significa que esses dados são impedidos de variar no intervalo entre $-\infty$ e $+\infty$ conforme requerido pelos métodos estatísticos clássicos.

O exemplo mais comum de dados composicionais é a distribuição do tamanho de partículas do solo, onde são expressas as frações areia, silte e argila. Quando expressos em termos de percentagens a soma dos valores dos três componentes é igual a 100, ou igual a 1000 quando expressos em termos de gramas por quilograma. Os exemplos se estendem para outras composições como os teores de cátions na solução do solo, ou as proporções dos diferentes usos da terra em uma região, a distribuição da porosidade do solo, entre outros.

Como os dados composicionais sempre somam uma constante eles apresentam características peculiares que os diferenciam dos demais tipos de dados. O principal deles é que o espaço amostral no qual os dados composicionais são comportados é o simplex que, no caso de composições de três componentes, é representado pelo diagrama ternário (o triângulo textural, por exemplo). Para composições de dois componentes o espaço amostral é uma reta e para composições de quatro componentes o espaço amostral é um tetraedro.

Todas essas características condicionam a maneira como as variáveis se relacionam entre si. E nesse caso a regra é que, obrigatoriamente, pelo menos uma covariância seja negativa e, consequentemente, um coeficiente de correlação seja negativo. Isso configura uma correlação espúria de tendência negativa, pois faz com que uma das características básicas do coeficiente de correlação não seja atingida, que é a sua liberdade de variar entre –1 e +1. Para o caso de uma composição com dois componentes o coeficiente de correlação será sempre igual a –1. Por esse motivo não é correto, em termos estatísticos, realizar análises de correlação de dados composicionais.

Outro problema associado às composições é o que se chama de incoerência subcomposicional, evidenciada pelo fato de que as relações de covariância entre os elementos em uma subcomposição não são as mesmas daquela observada entre os elementos da composição completa. Além disso, não há relação entre a estrutura da covariância da composição completa e da subcomposição. Para exemplificar o problema da incoerência subcomposicional imaginemos uma amostra de solo analisada em dois momentos (Figura 3). No primeiro momento tratamos a amostra de solo como contendo quatro componentes: areia, silte, argila e água, representada por $\begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix}$. Já no segundo momento decidimos secar a amostra e, assim, obtemos uma subcomposição com apenas três componentes: areia, silte e argila, representadas por $\begin{bmatrix} s_1 & s_2 & s_3 \end{bmatrix}$.

Composição completa $[x_1 \quad x_2 \quad x_3 \quad x_4]$	Subcomposição $[s_1 \quad s_2 \quad s_3]$
$[0{,}1 \quad 0{,}2 \quad 0{,}1 \quad 0{,}6]$	$[0{,}250 \quad 0{,}500 \quad 0{,}250]$
$[0{,}2 \quad 0{,}1 \quad 0{,}1 \quad 0{,}6]$	$[0{,}500 \quad 0{,}250 \quad 0{,}250]$
$[0{,}3 \quad 0{,}3 \quad 0{,}2 \quad 0{,}2]$	$[0{,}375 \quad 0{,}375 \quad 0{,}250]$

Figura 3 – Exemplo de composição completa contendo quatro componentes (areia, silte, argila e água) e uma subcomposição com apenas três componentes (areia, silte e argila). Adaptado de Aitchison (2003).

A razão entre os componentes, tanto na composição completa como na subcomposição, é a mesma:

$$\frac{x_1}{x_2} = \frac{s_1}{s_2} \tag{3}$$

Contudo, enquanto no primeiro momento obtemos uma correlação de 0.5 entre o teor de areia e silte, no segundo momento obtemos uma correlação de -1 entre as mesmas variáveis. A consequência dessa propriedade é que não podemos analisar dados composicionais em sua forma original através de métodos estatísticos baseados nas matrizes de variância-covariância ou na matriz de correlações. Isso inclui métodos como a análise fatorial, a análise discriminante e a análise de componentes principais.

Mais problemas surgem quando nosso objetivo é estimar a proporção de um dos componentes da composição de maneira isolada. É o caso da análise de regressão, onde obtemos três modelos diferentes para estimar, por exemplo, as proporções de areia, silte e argila em amostras de solo. Contudo, se somarmos os valores estimados pelos três modelos para cada um dos componentes veremos que o valor obtido não será uma constante, mas sempre menor ou maior. O mesmo raciocínio pode ser feito para estimativas espaciais, como por krigagem ou outros métodos de interpolação.

2.2.2 Tratamento de dados composicionais

Apesar dos apelos feitos por Aitchison sobre a necessidade de utilizar uma abordagem diferente para a análise de dados composicionais, poucos foram os resultados positivos. Boa parte dos pesquisadores reluta em utilizar uma abordagem diferente porque os métodos estatísticos clássicos ainda fornecem resultados que permitem uma interpretação (do ponto de vista geológico, pedológico, etc.) aparentemente coerente. Contudo, mesmo que isso seja verdade, os resultados são inválidos, pois os métodos são, comprovadamente, impróprios para a análise dos dados.

Como forma de superar o problema da análise estatística dos dados composicionais, Aitchison desenvolveu a teoria das log-razões. O autor observou que a tendência negativa na estrutura da covariância de dados composicionais é uma consequência das bases euclidianas da análise estatística clássica, onde a escala é absoluta e não relativa. Assim, o problema dos dados composicionais nada mais é do que um problema de escala. Como esses dados apresentam informações relativas das partes, não o seu valor absoluto, então a informação fornecida é essencialmente sobre a razão dos componentes. Mas como razões são difíceis de manejar tanto matemática como estatisticamente, uma vez que

$$\mathrm{var}\left(\frac{x_i}{x_j}\right) \neq \mathrm{var}\left(\frac{x_j}{x_i}\right) \tag{4}$$

onde var é a variância e x_i e x_j são dois vetores de dados, Aitchison propôs o uso de log-razões, uma vez que

$$\mathrm{var}\left\{\ln\left(\frac{x_i}{x_j}\right)\right\} = \mathrm{var}\left\{\ln\left(\frac{x_j}{x_i}\right)\right\} \tag{5}$$

onde ln é o logaritmo natural ou neperiano, de base *e*, onde *e* é um número irracional aproximadamente igual a 2,718281828459045, o número de Euler. Além disso, há uma correspondência biunívoca entre composições e um conjunto inteiro de log-razões:

$$\left[y_i \quad \cdots \quad y_{D-1}\right] = \left[\ln\left(\frac{x_i}{x_D}\right) \quad \cdots \quad \ln\left(\frac{x_{D-1}}{x_D}\right)\right], \text{ para todo } i = 1, 2, \cdots, D-1. \tag{6}$$

com inversa

$$\left[x_i \quad \cdots \quad x_D\right] = \left[\exp(y_i) \quad \cdots \quad \exp(y_{D-1}) \quad 1\right] / \left\{\exp(y_i) + \cdots + \exp(y_{D-1}) + 1\right\}. \tag{7}$$

Tal operação (Equação 6) transforma os dados composicionais do espaço fechado (o simplex) para o espaço real, ou seja, com variação entre $-\infty$ e $+\infty$, alterando a sua escala, que passa a ser absoluta. Assim, podemos aplicar qualquer método estatístico convencional aos dados e o resultado transformado para a escala original (relativa).

2.2.3 Log-razões

Aitchison (1982) apresentou duas transformações possíveis para dados composicionais: a log-razão aditiva (LRA) e a log-razão centralizada (LRC). Outras transformações foram desenvolvidas por outros autores, mas essas duas continuam sendo as mais utilizadas e de mais fácil compreensão. A principal diferença entre a transformação LRA e LRC é que enquanto a primeira resulta em um vetor com dois componentes, a segunda resulta em um vetor com três componentes. Os valores transformados passam, então, a ser chamados de coordenadas ou coeficientes.

Os coeficientes da LRA são os mais simples de obter e interpretar, onde os D–1 componentes são divididos pelo componente restante (geralmente o último) e, ao valor obtido, aplica-se o logaritmo natural. Esses coeficientes são geralmente utilizados em modelos de

estimativa como regressões lineares e interpolação em análise geoestatística como feito por Odeh et al. (2003), Lark e Bishop (2007) e Rawlins et al. (2009). A sua obtenção é feita conforme mostro na Figura 4 (considere-se uma composição x com D = 3 componentes expressos proporcionalmente).

$$x = [0{,}80 \quad 0{,}15 \quad 0{,}05]$$

$$LRA(x) = \left[\ln\frac{0{,}80}{0{,}05} \quad \ln\frac{0{,}15}{0{,}05} \right] = [\ln(16) \quad \ln(3)] = [2{,}77 \quad 1{,}10]$$

Figura 4 – Transformação de um vetor de dados x através da log-razão aditiva (LRA). Adaptado de Pawlowsky-Glahn e Egozcue (2006).

A transformação para a escala original é feita da seguinte maneira (Figura 5):

$$x = \left[\frac{\exp(2{,}77)}{\exp(2{,}77)+\exp(1{,}10)+1} \quad \frac{\exp(1{,}10)}{\exp(2{,}77)+\exp(1{,}10)+1} \quad \frac{1}{\exp(2{,}77)+\exp(1{,}10)+1} \right]$$

Figura 5 – Transformação de log-razões aditivas para a escala original. Adaptado de Pawlowsky-Glahn e Egozcue (2006).

Na obtenção dos coeficientes LRC dividimos os componentes pela média geométrica e, ao valor obtido, aplicamos o logaritmo natural. Esses coeficientes são mais difíceis de interpretar, mas tem a vantagem de poderem ser utilizados em algumas análises em que os coeficientes LRA não são adequados, como a análise log-contraste de componentes principais. A sua obtenção é feita da seguinte maneira (considere a composição mostrada acima) (Figura 6):

$$LRC(x) = \left[\ln \frac{0{,}80}{(0{,}80 \times 0{,}15 \times 0{,}05)^{1/3}} \quad \ln \frac{0{,}15}{(0{,}80 \times 0{,}15 \times 0{,}05)^{1/3}} \quad \ln \frac{0{,}05}{(0{,}80 \times 0{,}15 \times 0{,}05)^{1/3}} \right]$$

$$LRC(x) = \left[\ln \frac{0{,}80}{0{,}18} \quad \ln \frac{0{,}15}{0{,}18} \quad \ln \frac{0{,}05}{0{,}18} \right] = \left[\ln(4{,}4) \quad \ln(0{,}83) \quad \ln(0{,}28) \right] = \left[1{,}48 \quad -0{,}19 \quad -1{,}29 \right]$$

Figura 6 – Transformação de um vetor de dados *x* através da log-razão centralizada (LRC). Adaptado de Pawlowsky-Glahn e Egozcue (2006).

A transformação para a escala original é feita da seguinte maneira (Figura 7):

$$x = \left[\frac{\exp(1{,}48)}{\exp(1{,}48) + \exp(-0{,}19) + \exp(-1{,}29)} \quad \frac{\exp(-0{,}19)}{\exp(1{,}48) + \exp(-0{,}19) + \exp(-1{,}29)} \right.$$

$$\left. \frac{\exp(1{,}48)}{\exp(1{,}48) + \exp(-0{,}19) + \exp(-1{,}29)} \right]$$

Figura 7 – Transformação de log-razões centralizadas para a escala original. Adaptado de Pawlowsky-Glahn e Egozcue (2006).

A operação de transformação utilizada em ambos os casos é conhecida como operação de fechamento, onde cada componente do vetor é dividido pela soma de todos os componentes do vetor, levando a escala do vetor para a soma constante de 1.

3 HIPÓTESES

1 – Apesar da complexidade (geológica, topográfica, pedológica) e instabilidade da região do Rebordo do Planalto, é possível construir funções de predição espacial de propriedades do solo a partir de atributos de terreno que apresentam desempenho considerado satisfatório, ou seja, possuam a capacidade de explicar mais de 50% da variância.

2 – A estratificação da região do Rebordo do Planalto em domínios fisiográficos mais homogêneos do ponto de vista geológico permite a construção de funções de predição espacial que apresentam desempenho superior às anteriores, ou seja, com capacidade de explicar aproximadamente 70% da variância.

4 OBJETIVOS

4.1 Objetivo geral

Construir funções de predição espacial, a partir de parâmetros de superfície, para predizer a distribuição do tamanho de partículas da camada superficial dos solos da região do Rebordo do Planalto do RS.

4.2 Objetivos específicos

Identificar os fatores e processos que exercem maior influência sobre a distribuição do tamanho de partículas dos solos da região do Rebordo do Planalto do RS.

Avaliar o efeito da estratificação da região do Rebordo do Planalto do RS em domínios geologicamente mais homogêneos sobre a capacidade preditiva das funções de predição espacial construídas.

Identificar aspectos que devam ser levados em consideração em futuros estudos de mapeamento preditivo de propriedades dos solos da região do Rebordo do Planalto do RS.

5 MATERIAL E MÉTODOS

5.1 Área de estudo

A presente dissertação de mestrado dá continuidade aos trabalhos de Miguel (2010) e Samuel-Rosa et al. (2011b) realizados na sub-bacia Menino Deus I da bacia de captação do reservatório da Companhia Riograndense de Saneamento (CORSAN), a qual faz parte da bacia hidrográfica do arroio Vacacaí-Mirim. Esses dois autores já fizeram uma caracterização bastante detalhada da sub-bacia Menino Deus I. Assim, as informações que apresento aqui constituem um resumo daquelas apresentadas por aqueles autores. Para obtenção de maiores informações recomendo a consulta dos trabalhos citados acima.

A sub-bacia Menino Deus I está localizada na região do Rebordo do Planalto do RS, entre os municípios de Itaara e Santa Maria, com coordenadas UTM centrais 22 J 229489 m (E) e 6718530 m (S). Com uma área total de 18,92 km², a sub-bacia Menino Deus I corresponde a aproximadamente 60% da bacia de captação do reservatório da CORSAN (Dias, 2003), que é responsável por 30% do abastecimento de água da cidade de Santa Maria. O clima local é classificado como Cfa (subtropical úmido sem estação seca definida), com temperatura média anual de 19,3ºC e precipitação média anual de 1708 mm bem distribuídos aos longo do ano (Maluf, 2000). O relevo varia entre plano e montanhoso com elevações variando entre 139 e 475 m (Figura 8).

A geologia da sub-bacia Menino Deus I é bastante complexa (Brasil, 1980; Maciel Filho et al., 1988; Sartori, 2009): em elevações superiores a ±350 m ocorre a Sequência Superior da Formação Serra Geral (rochas ígneas – riolito-riodacito) e em elevações entre ±200 m e ±350 m ocorre a Sequência Inferior da Formação Serra Geral (rochas ígneas – basalto-andesito) e no seu interior ou abaixo dela a Formação Botucatu (rochas sedimentares – arenito eólico). Em elevações abaixo de ±200 m ocorre a Formação Caturrita (rochas sedimentares – arenito fluvial). Durante expedições de campo verificamos que existem depósitos coluviais de material proveniente de ambas as Formações Serra Geral e Botucatu em elevações entre ±200 e ±300 m. Ainda, depósitos coluviais de ambas as Formações Botucatu e Caturrita ocorrem em elevações abaixo de ±200 m. E, próximo às drenagens, podem ser encontrados depósitos fluviais recentes. Tais informações não são apresentadas nos

mapas geológicos disponíveis. Além disso, essas expedições de campo mostraram que os mapas têm sérias limitações em representar a variação dos materiais de origem na paisagem na escala requerida para o desenvolvimento desse estudo.

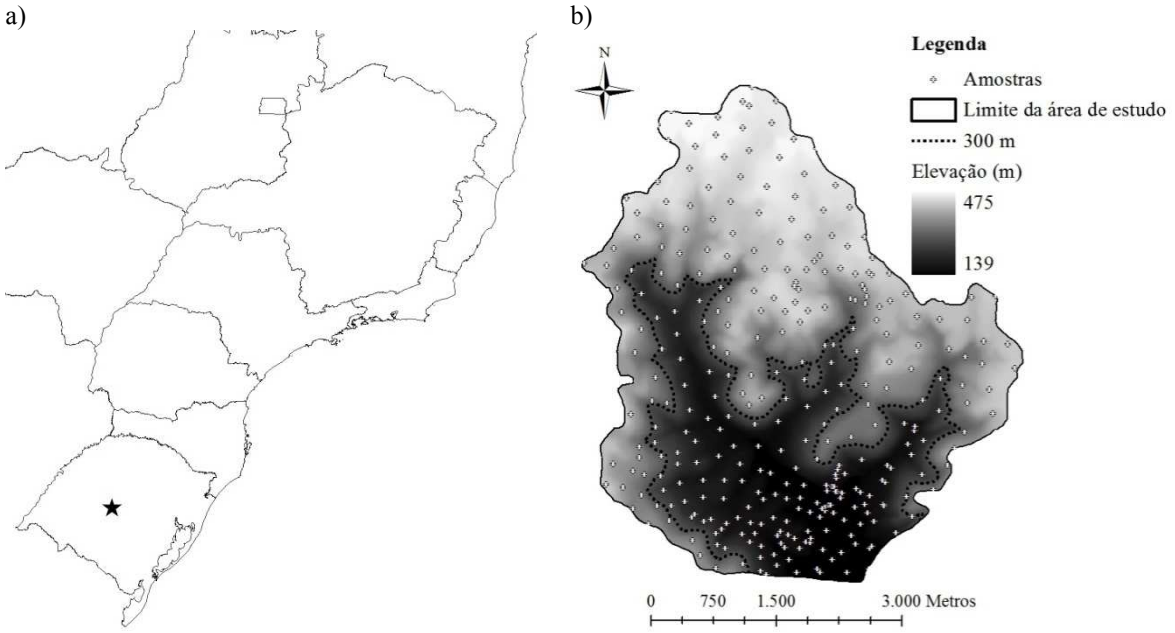


Figura 8 – (a) Localização da bacia de captação do reservatório da CORSAN estudada no presente trabalho e (b) representação da sub-bacia Menino Deus I com seu modelo digital de elevação, os pontos amostrados e a delimitação de dois domínios fisiográficos (superior – área mais clara, com solos derivados de rochas ígneas; e inferior – área mais escura, com solos derivados de rochas sedimentares) na cota de 300 m (linha pontilhada).

Predominam na área de estudo Neossolos Litólicos (mais de 50% da área), Argissolos Bruno - Acinzentados (14% da área), Argissolos Vermelhos (12% da área), e associações Cambissolo - Neossolo (14% da área) (Miguel, 2010). As áreas de floresta ocupam mais da metade da área, seguida pelas áreas de campo nativo, capoeira, lavoura, silvicultura, áreas urbanizadas e corpos d'água artificiais (Samuel-Rosa et al., 2011b).

Devido à complexidade geomórfica e pedológica da área de estudo, dividi a mesma em dois domínios fisiográficos mais homogêneos. Essa divisão foi feita com base no conhecimento de campo da variação do material de origem dos solos, o qual deve possuir forte influência sobre a distribuição do tamanho de partículas do solo. Assim, adotei a cota de 300 m (Figura 8) como divisão entre os dois domínios fisiográficos mais homogêneos. Em elevações acima de 300 m predominam solos derivados de rochas ígneas (Neossolos

Litólicos, Argissolos Vermelhos e Cambissolos), enquanto em elevações abaixo de 300 m predominam solos derivados de rochas sedimentares (Neossolos Litólicos, Argissolos Bruno-Acinzentados e Cambissolos).

5.2 Amostragem e análise dos solos

Os dados que utilizo no presente estudo são provenientes das amostras de solo e outras informações ambientais coletadas por Miguel (2010) e Samuel-Rosa et al. (2011b) quando da realização do levantamento de solos e de uso da terra da área de estudo. Esses pesquisadores coletaram trezentas e trinta e nove pontos amostrais (Figura 8). Devido à complexidade geomórfica da área, a escassez de recursos financeiros e de força de trabalho, além da ausência de informações a respeito da relação das propriedades do solo com outras variáveis ambientais, os pontos amostrais foram selecionados intencionalmente (*purposive sampling* (McKenzie et al., 2008)). Assim, o principal critério utilizado para seleção dos pontos amostrais foi a necessidade de amostrar feições geomórficas, manchas de solo e usos da terra representativos da área, utilizando como base o conhecimento pedológico dos pesquisadores. Uma das preocupações foi a necessidade de obtenção de uma cobertura amostral que fosse capaz de capturar a maior parte da variabilidade existente na área de estudo. Mas como diversos locais apresentaram sérias limitações ao acesso, os pontos amostrais acabaram ficando restritos às áreas onde o acesso era facilitado. A localização dos pontos amostrais no campo foi realizada utilizando as imagens de satélite disponibilizadas pelo Google Earth®. Através desse procedimento de coleta obteve-se uma densidade amostral de aproximadamente 18 pontos por km^2. A distância média mínima de separação entre dois pontos é de 181 m, variando de 18 a 328 m, com um desvio padrão de 80 m.

Em cada ponto amostral os pesquisadores definiram uma área de aproximadamente 100 m² no interior da qual abriram três trincheiras para coleta das amostras de solo. Essas amostras foram coletadas na camada superficial do solo de 0 a 20 cm ou do horizonte A inteiro quando o solo possuía espessura inferior a 20 cm. As três amostras foram utilizadas para produzir uma amostra composta que foi utilizada para as análises laboratoriais. Os pontos amostrais foram georreferenciados no campo utilizando um aparelho de GPS quando possível. Nas situações em que o sinal do aparelho receptor era comprometido, como no interior de florestas e próximo a encostas, o georeferrenciamento foi realizado diretamente na

tela do computador utilizando as imagens de satélite disponibilizadas pelo Google Earth®. O material de origem do solo foi identificado quanto ao tipo (ígneo ou sedimentar) com base nos mapas geológicos e nas características do solo e do ambiente adjacente (pedregosidade e rochosidade).

Em laboratório as amostras foram secas ao ar, destorroadas e passadas em peneira com malha de 2 mm. A distribuição do tamanho de partículas foi determinada através do método da pipeta depois da remoção da matéria orgânica com peróxido de hidrogênio (H_2O_2) (30 % v/v) naquelas amostras contendo mais que 5% de matéria orgânica (Embrapa, 1997). Hidróxido de sódio (NaOH) 1 mol L^{-1} foi utilizado como agente dispersante na análise. A Figura **9** mostra a distribuição do tamanho de partículas das amostras de solo de acordo com o material de origem dos solos (figura produzida utilizando o comando `TT.plot()` implementado no pacote *soiltexture* (Moyes e Shangguan, 2011) no ambiente R (R Development Core Team, 2011)).

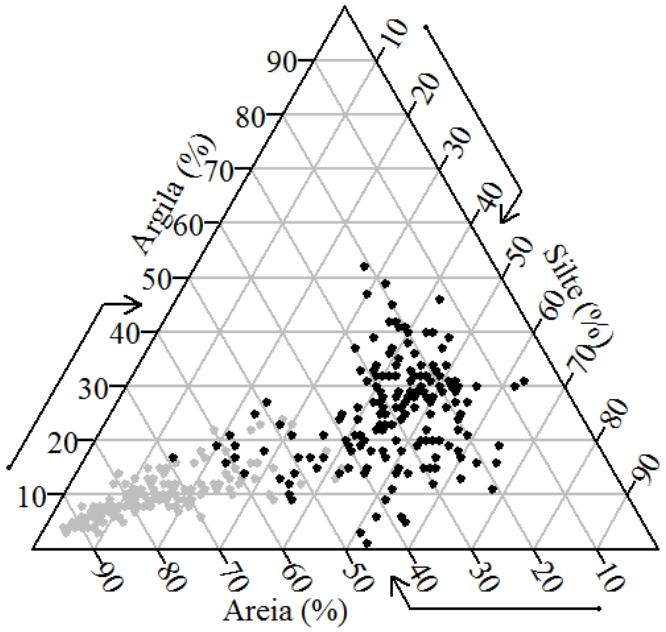


Figura 9 - Distribuição do tamanho de partículas das amostras de solo coletadas e a sua relação com o material de origem dos solos (preto – rochas ígneas, cinza – rochas sedimentares) (n = 339).

5.3 Modelo digital de elevação

O modelo digital de elevação (MDE) utilizado para derivar os atributos de terreno para construir as FPESe foi obtido a partir da interpolação das curvas de nível das cartas planialtimétricas Santa Maria (SH.22-V-C-IV/1 - SE) e Santa Maria (SH.22-V-C-IV-1 - NE) da Diretoria do Serviço Geográfico do Exército brasileiro (DSG) publicadas na escala de 1:25.000. A opção por não utilizar o MDE do Shuttle Radar Topography Mission (SRTM) se deve ao fato de que a sua resolução (90 m) não é adequada para atender aos objetivos do presente trabalho. Essa inadequação se deve ao fato de que quando estivermos estudando a variação espacial de propriedades do solo a resolução do MDE deve ficar entre 5 e 50 m (Hutchinson e Gallant, 2000). De fato, MDEs com resolução superior a 50 m não são visualmente capazes de representar as feições geomorfológicas observadas na área de estudo, causando a suavização da superfície.

A definição de um intervalo ótimo de resolução do MDE implica assumir que não existe um valor ótimo de resolução do MDE. Por esse motivo gerei quatro MDEs com resoluções de 5, 10, 20 e 40 m, ou seja, dentro do intervalo definido por Hutchinson e Gallant (2000). A interpolação dos MDEs foi realizada no ArcGIS 9.3 utilizando o comando *Topo to raster* implementado na caixa de ferramentas *3D Analyst*. *Topo to raster* constitui um método de interpolação baseado em uma antiga versão do programa ANUDEM desenvolvido por Hutchinson (1989). Ele incorpora um algoritmo que garante a criação de um MDE hidrologicamente correto e, portanto, sem a presença de depressões espúrias (ESRI, 2009). Isso significa que há uma coincidência acentuada entre a drenagem derivada numericamente e a hidrografia real.

Para seleção da resolução mais adequada utilizei o procedimento descrito por Hutchinson (1996). O método é baseado na variação da média quadrática da declividade com o aumento da resolução do MDE. A resolução mais adequada é determinada pelo refinamento da resolução do MDE até que a média quadrática da declividade estabilize (Hutchinson e Gallant, 2000). A teoria que dá suporte ao método desenvolvido por Hutchinson (1996) é a seguinte: em resoluções mais grosseiras, vários pontos do conjunto de dados base (as curvas de nível digitalizadas de cartas topográficas no presente estudo) são alocados em um mesmo pixel, o que leva a estimativa de um valor médio para o pixel. Isso resulta na suavização do MDE ajustado quando comparado a superfície real. Em resoluções mais finas esse efeito é minimizado, levando a estabilização das declividades do MDE ajustado. Quando

refinamentos subsequentes da resolução do MDE não produzem mais mudanças significativas nas declividades toda a informação contida no conjunto de dados base (as curvas de nível) foi extraída.

A média quadrática ($\bar{x}_q$) da declividade é calculada por

$$\bar{x}_q = \sqrt{\frac{x_i^2 + \cdots + x_n^2}{n}} \tag{8}$$

onde x_i é o valor da declividade no i-ésimo pixel com $i = 1, 2, ..., n$.

A Figura 10 mostra a variação da média quadrática da declividade com a variação da resolução do MDE. A curva mostra um significativo achatamento a partir da resolução de 10 m (11,4%) em direção a resolução de 5 m (11,6%). Isso sugere que o MDE com resolução ao redor de 10 m é o mais adequado para derivar os atributos de terreno. A partir da resolução de 5 m a curva deve ficar ainda mais achatada. Contudo, o uso de resoluções muito finas é indesejável por aumentar em demasia o tempo de processamento dos dados.

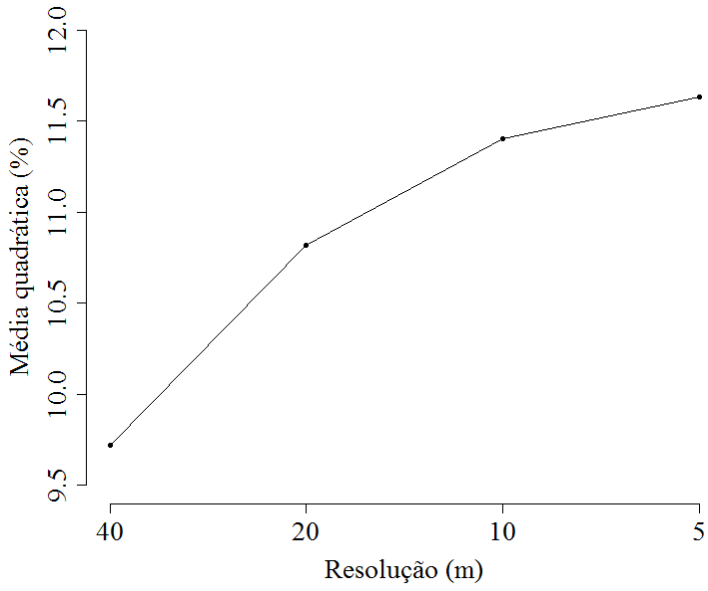


Figura 10 – Variação da média quadrática da declividade em função da resolução do MDE.

5.4 Variáveis preditoras

Apenas atributos de terreno foram utilizados como variáveis preditoras na construção das FPESe. Esses atributos foram derivados do MDE com resolução de 10 m utilizando o SAGA GIS (SAGA GIS, 2010). Quinze atributos de terreno (dez atributos primários e cinco atributos secundários) foram selecionados de acordo com (1) sua relação com os processos que têm influência sobre a distribuição do tamanho de partículas do solo e (2) seu uso comum na literatura. A seguir apresento os atributos de terreno utilizados, seu significado e importância, além dos procedimentos e equações para sua obtenção. Forneço também hiperlinks para acesso aos códigos-fonte do SAGA GIS.

5.4.1 Atributos primários

a) Área de Contribuição (inglês – Catchment Area, unidade – metro^2): atributo que representa a área acima da célula em questão que contribui para o fluxo superficial que chega até aquela célula. Está relacionado ao volume de enxurrada que chega até uma determinada célula. Foi calculado utilizando algoritmos de direção de fluxo múltiplo (Freeman, 1991; Quinn et al., 1991), os quais consideram que o fluxo entre as células não é unidirecional. Assim, "uma célula pode receber o fluxo proveniente de várias células e transferir o fluxo acumulado para várias outras células" (Minella et al., 2010).
b) Comprimento do Declive (inglês – Slope Length, unidade – metro): atributo relacionado à aceleração dos fluxos superficiais e as taxas de erosão e, portanto, ao Fator LS. Para sua obtenção parte-se das células de maior elevação em direção para as células de menor elevação, considerando-se a declividade de cada uma. Se a declividade da célula vizinha for maior do que o dobro da declividade da célula em questão, o comprimento do declive é a soma do comprimento da célula em questão e da célula vizinha. O algoritmo pára a acumulação de comprimento se a declividade da próxima célula vizinha for inferior a declividade da célula anterior. Se nenhuma das oito células vizinhas possui declividade superior dobro da declividade da célula em questão o CD é igual a zero.

c) Curvatura (inglês – Curvature, unidade – metro^{-1}): atributo que representa a combinação da curvatura plana e de perfil. Valores positivos indicam que a superfície é convexa para cima da célula em questão, enquanto valores negativos indicam que a superfície é côncava para cima da célula em questão. Um valor de zero indica que a superfície é plana. Ao considerar ambas as curvaturas plana e de perfil é possível ter um entendimento melhor dos fluxos superficiais. Foi calculado utilizando o método de Zevenbergen e Thorne (1987).

d) Curvatura de Perfil (inglês – Profile Curvature, unidade – metro^{-1}): atributo que representa a primeira derivada da declividade. Valores positivos descrevem curvaturas convexas, enquanto valores negativos descrevem curvaturas côncavas (Olaya, 2004). Possui influência sobre a velocidade do fluxo superficial, a taxa de erosão/deposição e a geomorfologia (Wilson e Gallant, 2000a). Foi calculado utilizando o método de Zevenbergen e Thorne (1987).

e) Curvatura Planar (inglês – Plan Curvature, unidade – metro^{-1}): atributo que representa a primeira derivada do aspecto. Valores positivos descrevem curvaturas convexas, enquanto valores negativos descrevem curvaturas côncavas (Olaya, 2004). Possui influência sobre a concentração (convergência) e dispersão (divergência) dos fluxos na paisagem, o que influencia diretamente o conteúdo de água no solo e as características do solo (Wilson e Gallant, 2000a). Foi calculado utilizando o método de Zevenbergen e Thorne (1987).

f) Declividade (inglês – Slope, unidade – graus): atributo que representa a primeira derivada da superfície de elevação no sentido do declive, perpendicular às curvas de nível. Expressa o gradiente ou taxa de mudança da elevação. Possui influência sobre a velocidade dos fluxos superficiais e subsuperficiais, o que influencia diretamente o conteúdo de água no solo, a taxa de erosão e a formação do solo (Wilson e Gallant, 2000a). Foi calculado utilizando o método de Zevenbergen e Thorne (1987).

g) Declividade Média da Área de Contribuição (inglês – Catchment Slope, unidade – graus): atributo que representa a declividade média da de todas as células que drenam para a célula em questão (Olaya, 2004). Está relacionado ao tempo de concentração, definido como o tempo necessário para que toda a AC contribua para o escoamento superficial que chega até a célula em questão. Assim, constitui um indicador da velocidade e potência dos fluxos superficiais (Olaya, 2009).

h) Elevação (inglês – Elevation, unidade – metro): atributo extraído diretamente do modelo digital de elevação (MDE). Representa a altitude da célula em questão em relação a um

plano de referência, geralmente o nível do mar. Possui influência sobre o clima, a vegetação e a energia potencial (Wilson e Gallant, 2000a).

i) Elevação Acima da Rede de Drenagem (inglês – Elevation Above Channel Network, unidade – metro): atributo que representa a distância vertical da célula em questão em relação à célula mais próxima localizada na rede de drenagem. Valores pequenos de EARD indicam locais em que o lençol freático pode estar mais próximo da superfície do solo, sendo caracterizadas como zonas de acumulação (Böhner et al., 2002). Assim sendo, a EARD está relacionada ao Índice de Umidade Topográfica (Olaya e Conrad, 2009). Já os valores intermediários indicam zonas de transferência de material, geralmente nos locais de maior declive (encostas), enquanto valores maiores indicam locais mais elevados da superfície geomórfica (possíveis zonas de perda de material) (Böhner et al., 2002).

j) Northerness (símbolo – NORT, inglês – Northerness, unidade - graus): atributo que indica a direção da vertente em relação ao norte. Obtido através da seguinte equação:

$$NORT = |180 - aspecto| \tag{9}$$

onde *aspecto* é o azimute da declividade expresso em graus no sentido horário a partir do norte, representando a primeira derivada da superfície de elevação ao longo do declive, paralelo às curvas de nível. O aspecto possui influência sobre a insolação, evapotranspiração, e a distribuição e abundância da flora e da fauna (Wilson e Gallant, 2000a). Mas como o aspecto constitui uma medida circular, seus valores não são adequados para comparação direta, sendo necessária a sua transformação para NORT (Roecker e Thompson, 2010). O aspecto foi calculado utilizando o método de Zevenbergen e Thorne (1987).

5.4.2 Atributos secundários

a) Fator LS (inglês – LS-Factor, unidade – adimensional): atributo equivalente ao fator topográfico da Equação Universal de Perda de Solo Revisada (RUSLE) que representa o efeito da topografia sobre a erosão (quanto maior o LS, maior o potencial erosivo), além

de caracterizar os processos de erosão e deposição (Moore et al., 1993). Na Equação Universal de Perda de Solo (EUPS) o LS é calculado utilizando equações que consideram uma vertente de relevo uniforme (encosta retilínea), tendo como referência a parcela padrão de 22,13 m. Mas em áreas de grande extensão e de relevo complexo o LS assume uma dimensão de área ou uma unidade hidrológica representativa da bacia (Minella et al., 2010). Nesse caso devem ser utilizadas equações que levem em consideração os fluxos divergentes e convergentes do escoamento superficial. Um dos métodos é aquele desenvolvido por (Moore et al., 1991), que incorpora a teoria da potência unitária do escoamento, segundo a qual a água na superfície do solo apresenta determinada energia capaz de desagregar e transportar partículas de solo quando estas se movem no sentido do declive (Minella et al., 2010). Assim, o LS é obtido através da seguinte equação:

$$LS = (n+1) \times \left(\frac{AC_s}{22{,}13} \right)^n \times \left(\frac{sen(\beta)}{0{,}0896} \right)^m \quad (10)$$

onde AC_s é a área de contribuição específica ($m^2\ m^{-1}$), β é a declividade (graus), n = 0,4 e m = 1,3. A área de contribuição específica é dada por:

$$AC_s = \frac{AC}{célula} \quad (11)$$

onde AC é a área de contribuição (m^2) e *célula* corresponde ao tamanho da célula ou pixel (m).

b) Índice de convergência (inglês – Convergence Index, unidade – porcentagem): atributo desenvolvido por Köethe e Lehmeier (1996) apud Conrad (1998) que integra a informação dos valores de curvatura e assim fornece uma maneira mais fácil de interpretar o comportamento do fluxo (Olaya, 2004). Esse índice é derivado dos desvios dos valores de aspecto de todas as células vizinhas em relação à célula central de uma janela de 3 por 3 células (Conrad, 1998; Olaya e Conrad, 2009). A soma das diferenças é expressa em termos de percentagens, onde +100% indica divergência

total, 0% indica uma superfície plana (todas as células vizinhas possuem aspecto paralelo) e –100% indica convergência total (Figura 11) (Conrad, 1998).

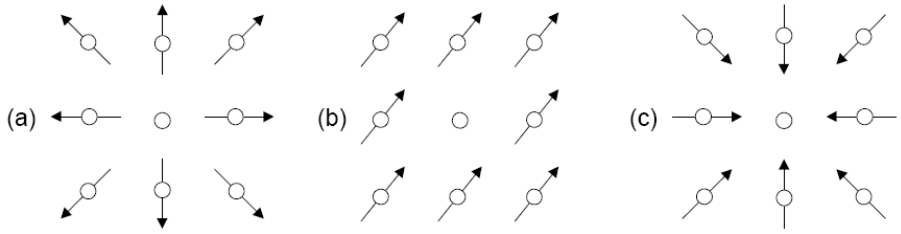


Figura 11 – O Índice de Convergência é calculado a partir do aspecto das oito células vizinhas. O exemplo mostra (a) divergência total, (b) superfície plana e (c) convergência total. Adaptado de Conrad (1998).

c) Índice de Potência de Escoamento (inglês – Stream Power Index, unidade – adimensional): atributo desenvolvido por Moore et al. (1988), trata-se de uma medida do potencial erosivo da enxurrada. Com o aumento da área de contribuição específica e da declividade, a quantidade de água que chega das áreas a montante e a velocidade do fluxo da água aumentam, levando ao consequente aumento da potencial de escoamento e de erosão (Gruber e Peckham, 2009). O IPE é obtido através da seguinte equação:

$$IPE = AC_s \times \tan(\beta) \qquad (12)$$

onde AC_s é a área de contribuição específica (m^2 m^{-1}) e β é a declividade (graus). O IPE pode ser utilizado para identificar os locais onde medidas de conservação do solo que reduzem os efeitos erosivos do escoamento concentrado podem ser adotadas (Moore et al., 1991).

d) Índice de Rugosidade da Superfície (inglês – Terrain Ruggedness Index, unidade – adimensional): atributo desenvolvido por Riley et al. (1999) que quantifica a heterogeneidade da superfície. Está relacionado aos processos de formação do solo, a geomorfologia e a distribuição da fauna e da flora. O índice é calculado a partir da soma da mudança de elevação entre uma célula e as suas oito células vizinhas (Figura 12).

-1,-1	0,-1	1,-1
-1,0	0,0	1,0
-1,1	0,1	1,1

Figura 12 – Representação de uma janela de 3 por 3 células de uma superfície de elevação para o cálculo do Índice de Rugosidade da Superfície. Adaptado de Riley et al. (1999).

Se cada quadrado na Figura 12 representa uma célula de uma superfície de elevação, então o IRS é dado pela seguinte equação:

$$IRS = y\left[\sum \left(x_{ij} - x_{00}\right)^2\right]^{\frac{1}{2}} \quad (13)$$

onde x_{ij} é a elevação de cada célula vizinha a célula (0,0). Os valores de IRS variam entre os seguintes limites: plano = 0 a 80 m; aproximadamente plano = 81 a 116 m; levemente rugoso = 117 a 161 m; rugosidade intermediária = 162 a 239 m; moderadamente rugoso = 240 a 497 m; altamente rugoso = 498 a 958 m; e, extremamente rugoso ≥ 958 m. Cada classe recebe um número inteiro como identificador.

e) Índice de Umidade Topográfica (inglês – Topographic Wetness Index, unidade – adimensional): atributo desenvolvido por Beven e Kirkby (1978) que descreve a tendência de uma célula acumular água (Gruber e Peckham, 2009). Assim, maiores valores de IUT indicam maior tendência de acumular água e, portanto, maior conteúdo de água no solo. O IUT é obtido através da seguinte equação:

$$IUT = \ln\left[\frac{AC_s}{\tan(\beta)}\right] \quad (14)$$

onde A_s é a área de contribuição específica (m^2 m^{-1}) e β é a declividade (graus). Contudo, essa equação assume condições estáveis e solos com propriedades uniformes, ou seja, a transmissividade é constante ao longo da bacia e igual à unidade (Wilson e Gallant, 2000a). Como a variação do componente topográfico é comumente maior do que a variação da transmissividade do solo (Wood et al., 1990), essa equação pode ser usada na maioria das superfícies geomórficas.

5.4.3 Processamento das variáveis preditoras

A informação contida nos 15 planos de informação contendo os atributos de terreno foi amostrada para os 339 pontos georreferenciados onde foram coletadas amostras de solo. Para isso utilizei o algoritmo do vizinho mais próximo (inglês – nearest neighbor) no SAGA GIS. Esse algoritmo amostra o valor da célula mais próxima de onde se encontra o ponto georeferenciado. Como utilizei um MDE com resolução de 10 m, os pontos de coleta acabam sendo localizados no interior de uma célula individual de 100 m². Assim, o algoritmo do vizinho mais próximo amostra o valor da célula do plano de informação em que está contido. Os dados amostrados dos atributos de terreno, doravante chamados variáveis preditoras, foram importados para o ambiente R onde estatísticas descritivas e histogramas de frequência foram obtidos utilizando as funções `descdist()` e `plotdist()` implementadas no pacote *fitdistrplus* (Delignette-Muller et al., 2010). Variáveis preditoras com coeficiente de assimetria superior a 0,50 foram transformadas para a escala logarítmica natural ou para a sua raiz quadrada para obter uma distribuição próxima da normal. Essas estatísticas também foram utilizadas para verificar a qualidade do conjunto de dados das variáveis preditoras.

Uma das maiores preocupações que tive durante o processamento das variáveis preditoras foi a identificação da ocorrência de multicolinearidade (também conhecida como redundância). A multicolinearidade representa o grau de explicação do efeito de uma variável preditora sobre a variável dependente pelas outras variáveis preditoras utilizadas na análise

(Hair et al., 2010). O seu efeito na análise é negativo, uma vez que o verdadeiro efeito das variáveis preditoras não pode ser determinado à medida que a multicolinearidade aumenta. Como consequência, pode-se realizar uma interpretação equivocada dos resultados da análise estatística. Assim, deve-se utilizar variáveis preditoras que possuam baixa multicolinearidade, mas que possuam elevada correlação com a variável dependente.

A estratégia mais simples de identificação da ocorrência de multicolinearidade é o uso da matriz de correlação linear das variáveis preditoras, da qual lanço mão no presente estudo. Para isso escolhe-se um valor elevado do coeficiente de correlação (Hair et al., 2010), o qual é estabelecido como o limite máximo aceitável de correlação entre duas variáveis preditoras a serem inseridas na análise. No meu caso decidi por utilizar o valor do coeficiente de correlação de 0,80 (note que um r = 0,80 indica que as variáveis "compartilham" 64% da variância → 0,80 × 0,80 = 0,64). Assim, uma das variáveis foi eliminada quando o coeficiente de correlação entre duas variáveis preditoras foi ≥ 0,80, mantendo-se aquela com maior relevância conceitual. Nesse caso, dei preferência ao uso de variáveis preditoras que estão relacionadas aos processos físicos que ocorrem ao longo da paisagem e que podem apresentar influência sobre a distribuição do tamanho de partículas do solo. Segundo Moore et al. (1993), são os atributos de terreno secundários aqueles capazes de desempenhar esse papel. Além disso, os atributos de terreno secundários já possuem, implicitamente, informações sobre os atributos de terreno primários, uma vez que são construídos a partir desses. Isso indica que o uso concomitante de atributos de terreno primários e secundários, construídos a partir daqueles, constitui a ocorrência de uma redundância teórica e, possivelmente, de multicolinearidade.

O segundo método estatístico que utilizei para identificação da ocorrência de multicolinearidade foi a análise de componentes principais (ten Caten et al, 2011b). Através do teste de esfericidade de Bartlett e dos testes de adequação amostral KMO (Kaiser-Meyer-Olkin) e MAS (Measure of Sample Adequacy) verifiquei a adequação dos dados a análise de componentes principais. O teste de esfericidade de Bartlett é utilizado para verificar a hipótese de que as variáveis não são correlacionadas entre si. Em outras palavras, isso significa que a matriz de correlação é uma matriz identidade (Figura 13), onde cada variável correlaciona-se perfeitamente consigo mesma (r = 1,0), mas não possui qualquer correção com as demais variáveis (r = 0,0). Na maioria dos casos existe correlação entre as variáveis e o resultado do teste é significativo. O teste de esfericidade de Bartlett foi executado através da função `cortest.bartlett()` implementada no pacote *psych* (Revelle, 2011).

$$\begin{bmatrix} 1{,}0 & 0{,}0 & 0{,}0 & 0{,}0 \\ 0{,}0 & 1{,}0 & 0{,}0 & 0{,}0 \\ 0{,}0 & 0{,}0 & 1{,}0 & 0{,}0 \\ 0{,}0 & 0{,}0 & 0{,}0 & 1{,}0 \end{bmatrix}_{4x4}$$

Figura 13 – Exemplo de uma matriz identidade representando a matriz de correlação linear de um conjunto de variáveis não correlacionadas entre si.

Já os testes MSA e KMO não fornecem uma estatística de probabilidade para aceitação ou rejeição da hipótese básica. Na verdade, os dois testes fornecem uma estatística a partir da qual se faz a inferência. De maneira geral, valores entre 0,5 e 1,0 são desejáveis, indicando que os dados são adequados à análise de componentes principais. Valores abaixo de 0,5 indicam que a análise de componentes principais pode não ser adequada para o conjunto de dados atual. Como forma de "adequar" o conjunto de dados à análise faz-se a remoção das variáveis que possuem um valor do índice MSA $\leq 0{,}5$, começando por aquela que possui o menor valor. A remoção deve ser feita de uma variável por vez, realizando-se o teste novamente após a remoção de cada variável. Os testes de adequação amostral foram realizados utilizando a função `kmo()` escrita por G. Jay Kerns.

A análise de componentes principais foi realizada com as variáveis preditoras que se mostraram consistentes nos testes KMO e MSA. Para isso utilizei os comandos `eigen()`, implementado no pacote *base*, e `princomp()`, implementado no pacote *stats* (R Development Core Team, 2011). A análise foi feita a partir da matriz de correlação das variáveis preditoras. Os autovalores foram convertidos para coeficientes de correlação (c_{ij}) entre os escores das componentes e as variáveis originais por

$$c_{ij} = a_{ij}\sqrt{\frac{v_j}{s_i^2}} \qquad (15)$$

onde a_{ij} é o i-ésimo elemento do j-ésimo autovetor, v_j é o j-ésimo autovalor, e $s^2{}_i$ é a variância da i-ésima variável original. Como a análise de componentes principais foi realizada

utilizando a matriz de correlação das variáveis preditoras, s^2_i = 1,0. Os coeficientes foram plotados em círculos unitários nos planos da primeira (CP 1), segunda (CP 2) e terceira (CP 3) dimensões para auxiliar na identificação da ocorrência de multicolinearidade. Nesse caso, a ocorrência de agrupamentos de variáveis preditoras foi interpretada como sendo resultado da ocorrência de multicolinearidade, o que significa que as variáveis no mesmo agrupamento são correlacionadas. Também utilizei a matriz de pesos das variáveis preditoras em cada componente principal extraída para avaliar a ocorrência de multicolinearidade. Variáveis preditoras com pesos elevados em uma mesma componente principal foram consideradas colineares. Em ambos os casos utilizei como critério de exclusão de variáveis a sua relevância conceitual, conforme já descrevi acima na análise de correlação linear.

De posse das variáveis selecionadas para construção das FPESe, calculei sua correlação com as frações de tamanho de partícula (argila, silte e areia). A significância das correlações foi testada utilizando o comando `r.test()` implementado no pacote *psych* (Revelle, 2011).

5.5 Construção e avaliação das FPESe

5.5.1 Ajuste dos modelos de regressão linear múltipla

As funções de predição espacial da distribuição do tamanho de partículas do solo foram construídas através do ajuste de modelos de regressão linear múltipla:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \qquad (16)$$

onde Y_i é o valor da variável dependente na i-ésima observação, β_0, β_1, β_2 ... β_{p-1} são parâmetros, X_{i1}, X_{i2} ... $X_{i,p-1}$ são constantes conhecidas, ou seja, o valor das variáveis preditoras na i-ésima observação, ε_i é o erro aleatório com média $E\{\varepsilon_i\} = 0$ e variância $\sigma^2\{\varepsilon_i\} = \sigma^2$, e i = 1, ... , n (Kutner et al., 2004).

O comando `lm()`, implementado no pacote *stats* (R Development Core Team, 2011), foi utilizado para realizar o ajuste dos modelos regressão linear múltipla. Mas antes do ajuste dos modelos de regressão linear múltipla os dados das três frações (areia, silte e argila) foram transformados em razões log-aditivas conforme preconizado por Aitchison (1982) (Equação (6)), resultando em duas variáveis dependentes: *ln*(*argila/areia*) e *ln*(*silte/areia*).

A partir da matriz de dados das n = 339 observações, contendo os valores das duas variáveis dependentes (*ln*(*argila/areia*) e *ln*(*silte/areia*)) e das quatro variáveis preditoras (CONV, ELEV, IPE e IUT), outras duas foram derivadas. A primeira delas contendo as observações localizadas em elevações < 300 m (domínio inferior), ou seja, onde predominam solos derivados de rochas sedimentares (Figura 8). O número de observações contido nesse conjunto de dados é igual a n = 165. A segunda matriz de dados derivada contém as observações localizadas em elevações ≥ 300 m (domínio superior), ou seja, onde predominam solos derivados de rochas ígneas (Figura 8). O número de observações contido nesse conjunto de dados é igual a n = 174.

Em seguida, amostrei aleatoriamente subconjuntos de observações (comando `sample()` implementado no pacote *base* (R Development Core Team, 2011)) das três matrizes de dados (n = 339; n = 165; n = 174). Da primeira matriz (contendo os dados de toda a área) amostrei n = 300 observações, enquanto que de cada uma das matrizes dos domínios fisiográficos amostrei n = 150 observações, a partir dos quais ajustei os modelos de regressão. Dois foram os objetivos ao realizar esse procedimento. O primeiro foi introduzir certo grau de aleatoriedade ao conjunto de dados. Isso é importante porque os pontos de amostragem foram selecionados com base no conhecimento tácito dos pedólogos que realizaram o levantamento dos solos e do uso da terra na área de estudo (Miguel, 2010; Samuel-Rosa et al., 2011b). O procedimento mais adequado seria uma amostragem probabilística, utilizando pontos de amostragem selecionados aleatoriamente (Webster e Oliver, 1990). Como isso não foi possível, decidi por realizar uma amostragem aleatória do conjunto de dados original. Fica, contudo, a necessidade de avaliar se esse procedimento teve algum efeito significativo sobre os resultados obtidos.

O segundo objetivo da obtenção de uma amostra aleatória do conjunto inicial de dados está relacionado ao procedimento de validação dos modelos de regressão ajustados e será exposto na seção seguinte que trata sobre esse assunto.

Para iniciar o algoritmo gerador de números aleatórios, utilizei como semente aleatória o número 123, definida através do comando `set.seed()` implementado no pacote *base* (R Development Core Team, 2011). Não há qualquer critério estatístico ou matemático que

fundamente a utilização do número 123 como semente aleatória. Sua escolha se deu por se tratar de uma sequência numérica de fácil memorização. O uso de uma semente aleatória garante que a mesma sequência de números aleatórios seja gerada sempre que o algoritmo gerador de números aleatórios for acionado. Assim, sempre que se desejar obter uma amostra com, por exemplo, 300 observações do conjunto inicial de 339 observações e utilizada como semente aleatória o número 123, exatamente as mesmas observações serão amostradas. Minha intenção ao definir essa estratégia é permitir a inteira reprodutibilidade do procedimento analítico que desenvolvo para ajustar os modelos de regressão. Contudo, permanece a necessidade de avaliar se o uso de diferentes conjuntos de dados teria efeito sobre os modelos de regressão ajustados.

Os três novos conjuntos de dados selecionados aleatoriamente foram utilizados para ajustar os modelos de regressão linear múltipla (FPESe) para estimar a distribuição do tamanho de partículas do solo em toda a área de estudo (n = 300) e nos domínios inferior (n = 150) e superior (n = 150). As observações restantes (aquelas que não foram amostradas através do procedimento descrito acima → 339 – 300 = 39; 165 – 150 = 15; 174 – 150 = 24) foram reservadas para uso durante o procedimento de validação dos modelos de regressão ajustados (descrito abaixo).

Durante a análise de regressão as variáveis preditoras foram submetidas ao procedimento stepwise bidirecional utilizando o comando `stepAIC()` implementado no pacote *MASS* (Venables e Ripley, 2002). O procedimento stepwise bidirecional permite a seleção das variáveis preditoras com maior contribuição para o modelo de regressão ajustado. Para isso, o critério utilizado para definir a melhor solução de regressão foi a minimização do Critério de Informação de Akaike dado por

$$AIC = -2 \times L + 2 \times p \tag{17}$$

onde AIC é o Critério de Informação de Akaike, L é a estatística log verossimilhança e p é o número de parâmetros utilizados no ajuste do modelo (Venables e Ripley, 2002). Como se pode ver na equação (17), o AIC constitui um índice que considera o número de parâmetros utilizados no ajuste do modelo de regressão e a qualidade do ajuste a ele associado. Assim, o procedimento *stepwise* bidirecional seleciona o modelo de regressão que inclui um determinado número de variáveis preditoras que resulta no menor valor de AIC. A seleção

ocorre da seguinte maneira: o algoritmo inicia o modelo sem nenhuma variável preditora; em seguida o algoritmo adiciona as variáveis uma a uma e seleciona aquela que resulta na maior minimização de *AIC*; a cada novo passo um processo de eliminação é realizado para remover as variáveis preditoras que não são mais capazes de melhorar o modelo através da minimização do *AIC*. As variáveis preditoras selecionadas através desse procedimento foram utilizadas no ajuste dos modelos de regressão linear.

5.5.2 Avaliação das FPESe

A qualidade do ajuste dos modelos de regressão linear múltipla (FPESe) foi avaliada graficamente (comando `plot()` implementado no pacote *graphics* (R Development Core Team, 2011)) utilizando os resíduos, os resíduos padronizados, a distância de Cook e a alavancagem.

Os resíduos (res_i) são dados por

$$res_i = \hat{y}_i - y_i \tag{18}$$

onde res_i são os resíduos, $\hat{y}_i$ é o valor predito na i-ésima observação e y_i é o valor medido na i-ésima observação, com $i = 1, 2, ..., n$ (Kutner et al., 2004).

Os resíduos padronizados (res_i^*) são dados por

$$res_i^* = \frac{res_i}{s_r} \tag{19}$$

onde res_i^* são os resíduos padronizados, res_i é o resíduo da i-ésima observação e s_r é o desvio padrão dos resíduos, com $i = 1, 2, ..., n$ (Hair et al., 2010). Os resíduos padronizados possuem média igual a 0 e desvio padrão igual a 1. Para amostras grandes ($n > 50$), os resíduos

padronizados seguem aproximadamente a distribuição *t*, o que implica no fato de que resíduos padronizados > 1,96 (valor *t* crítico no nível de confiança 0,05) podem ser considerados estatisticamente significantes (Hair et al., 2010). Isso significa que observações com resíduos padronizados > 1,96 podem ser consideradas atípicas (*outliers*).

A distância de Cook é dada por

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{y}_j - \hat{y}_{j(i)} \right)^2}{p \times QM_{erro}} \tag{20}$$

onde D_i é a distância de Cook, $\hat{y}_j$ é o valor predito na *j*-ésima observação quando todas as *n* observações são utilizadas para ajustar o modelo de regressão, $\hat{y}_{j(i)}$ é o valor predito na *j*-ésima observação quando a i-ésima observação é eliminada do ajuste do modelo de regressão, *p* é o número de graus de liberdade do modelo de regressão e QM_{erro} é o quadrado médio do erro (o denominador serve como medida de padronização) (Kutner et al., 2004). Assim, D_i consiste em uma medida agregada de influência das observações, ou seja, ela avaliar a influência da *i*-ésima observação sobre todos os *n* valores preditos. A avaliação de D_i foi realizada com base na recomendação de McDonald (2002), para quem uma observação é influencial ($n > 15$) se $D_i > 0{,}70$ para $p = 2$ (uma variável preditora), $D_i > 0{,}80$ para $p = 3$ (duas variáveis preditoras) e $D_i > 0{,}85$ para $p > 3$ (mais de duas variáveis preditoras). Outra possibilidade seria a avaliação das distâncias relativas de Cook, ou seja, verificar a existência de observações que apresentem D_i muito discrepante das demais. Essas observações seriam, então, consideradas influenciais. Contudo, se o maior D_i for substancialmente inferior a 1,0 a eliminação da observação correspondente não resultará em alteração significativa da estimativa dos parâmetros do modelo ajustado (Weisberg, 2005).

A alavancagem (h_{ii}) é obtida da diagonal da matriz de projeção e indica a magnitude do impacto que uma observação possui sobre os resultados do modelo de regressão ajustado devido às suas diferenças em relação às outras observações no que diz respeito a uma ou mais variáveis preditoras (Hair et al., 2010). Seu valor vaia entre 0 e 1, e o valor médio é igual a *p/n*, onde *p* é o número de parâmetros (variáveis preditoras mais a constante) no modelo de regressão e *n* é o número de observações utilizadas para ajustar o modelo de regressão. Hair et al. (2010) sugerem que para modelos de regressão ajustados com menos de dez parâmetros

seja utilizado como limite máximo aceitável de alavancagem o valor de três vezes a média ($3p/n$).

Em cada gráfico produzido utilizando res_i, res_i^*, D_i e h_{ii}, as seis observações que se apresentaram atípicas ou mais influentes foram identificadas. Sua exclusão do conjunto de dados utilizado para realizar o ajuste dos modelos de regressão foi feito com base nos critérios expostos acima. Quando mantidas, busquei identificar os motivos pelos quais tais observações se mostraram atípicas ou influenciais, sempre justificando sua manutenção.

Depois de ajustados os modelos de regressão, utilizei a análise de variância (ANOVA) (comando `anova.lm()` implementado no pacote *stats* (R Development Core Team, 2011)) para realizar a verificação final da adequação do ajuste dos modelos de regressão aos conjuntos de dados. O comando `anova.lm()` fornece a soma de quadrados extra (SQ_{extra}) de cada variável preditora, dado que a(s) variável(is) preditora(s) anterior(s) já está(ao) no modelo. A SQ_{extra} fornece a possibilidade de verificar o acréscimo na soma de quadrados da regressão ($SQ_{regressão}$) quando uma ou mais variáveis preditoras são adicionadas ao modelo de regressão. Além disso, o comando `anova.lm()` fornece os graus de liberdade associados a cada variável preditora. Ao somar as SQ_{extra} e os graus de liberdade das variáveis preditoras obtém-se a $SQ_{regressão}$ e os graus de liberdade a ela associados. Ao dividir a $SQ_{regressão}$ pelos seus graus de liberdade obtém-se o quadrado médio da regressão ($QM_{regressão}$). A soma de quadrados total (SQ_{total}) é obtida pela soma de $SQ_{regressão}$ e $SQ_{resíduos}$. Com os valores das somas de quadrados foi possível determinar a partição da variância total dos dados, ou seja, a contribuição de cada variável preditora para a explicação da variância (EV) do conjunto de dados. Assim

$$EV(\%) = \frac{SQ_{total}}{SQ_{extra\,(i)}} \times 100 \qquad (21)$$

onde EV(%) é percentual da variância contida nos dados explicada pela variável preditora p, SQ_{total} é a soma de quadrados total, calculada conforme descrito acima, que estima a variância, e $SQ_{extra\,(i)}$ é a soma de quadrados extra associado a i-ésima variável preditora, com $i = 1, 2, ..., p$.

A matriz variância-covariância (MVC) dos parâmetros ajustados (comando `vcov()` implementado no pacote *stats* (R Development Core Team, 2011)) e o fator de inflação da

variância (*FIV*) associado a cada variável preditora utilizada no ajuste dos modelos de regressão (comando `vif()` implementado no pacote *HH* (Heiberger, 2011)) foram utilizados para avaliar a ocorrência de multicolinearidade nos modelos de regressão ajustados. A ocorrência de multicolinearidade provoca o aumento da variância dos parâmetros ajustados, a qual é encontrada na diagonal da *MVC* (Hair et al., 2010). Assim, a variável preditora cujo parâmetro ajustado no modelo de regressão seja elevado (comparativamente aos demais) está correlacionada às demais variáveis preditoras. O *FIV* auxilia na identificação da ocorrência de multicolinearidade, sendo dado por

$$FIV_i = \frac{1}{1 - R_i^2} \tag{22}$$

onde FIV_i é o fator de inflação da variância da *i*-ésima variável preditora utilizada no ajuste do modelo de regressão e R_i^2 é o coeficiente de determinação do modelo de regressão ajustado para estimar a *i*-ésima variável preditora a partir das demais variáveis preditoras, com $i = 1, 2, ..., p$ (Hair et al., 2010). Assim, quanto maior o *FIV*, maior é a multicolinearidade. O valor de *FIV* mais comumente adotado como limite máximo aceitável é $FIV = 10$ (Hair et al., 2010).

Tendo os modelos de regressão sido ajustados e definidos para os três conjuntos de dados ($n = 300$; $n = 150$; $n = 150$), transformei os valores preditos das log-razões aditivas *ln*(*argila/areia*) e *ln*(*silte/areia*) para a escala original (argila, silte e areia) conforme preconizado por Aitchison (1982) (Equação (7)). Esses valores foram então comparados graficamente (comando `plot()` implementado no pacote *graphics* (R Development Core Team, 2011)) aos valores medidos nas observações que utilizei para ajustar os modelos de regressão. Em seguida, utilizei as FPESe construídas para estimar a distribuição do tamanho de partículas das $n = 339$ observações. Para isso utilizei o comando `predict()` implementado no pacote *stat* (R Development Core Team, 2011). De posse dos valores preditos e dos valores medidos, calculei os res_i de cada uma das frações de tamanho de partícula (argila, silte e areia) (Equação (18)). Esses res_i foram submetidos à análise variográfica para identificar a estrutura de sua variação espacial e assim verificar a possibilidade de gerar mapas de sua distribuição espacial via krigagem. Para isso utilizei o semivariograma, que permite representar quantitativamente a variação de um fenômeno regionalizado no espaço (Huijbregts, 1975). O nível de dependência espacial de dois pontos

no espaço é dado pelo variograma ($2\gamma(\boldsymbol{h})$) (Camargo, 1997), definido como a esperança matemática do quadrado da diferença entre os valores de observações no espaço, separadas pelo vetor distância $\boldsymbol{h}$, ou seja,

$$2\gamma(\boldsymbol{h}) = E\left\{\left[Z(\boldsymbol{x}) - Z[\boldsymbol{x}+\boldsymbol{h}]\right]^2\right\} = Var\left[Z(\boldsymbol{x}) - Z(\boldsymbol{x}+\boldsymbol{h})\right]. \tag{23}$$

Utilizando uma amostra $Z(\boldsymbol{x}_i)$, com $i = 1, 2, ..., n$, o variograma pode ser estimado por

$$2\hat{\gamma}(\boldsymbol{h}) = \frac{1}{N(\boldsymbol{h})}\sum_{i=1}^{n(\boldsymbol{h})}\left[z(\boldsymbol{x}_i) - z(\boldsymbol{x}_i+\boldsymbol{h})\right]^2, \tag{24}$$

onde $2\hat{\gamma}(\boldsymbol{h})$ é o variograma estimado, $N(\boldsymbol{h})$ é o número de pares valores medidos, $z(x_i)$ e $z(x_i + \boldsymbol{h})$, separados por um vetor distância $\boldsymbol{h}$, e $z(x_i)$ e $z(x_i + \boldsymbol{h})$ são os valores da i-ésima observação da variável regionalizada, coletados nos pontos $\boldsymbol{x}_i$ e $\boldsymbol{x}_i + \boldsymbol{h}$, com $i = 1, 2, ..., n$, separados pelo vetor $\boldsymbol{h}$. Assim sendo, o semivariograma é dado por

$$\gamma(\boldsymbol{h}) = \frac{1}{2}E\left\{\left[Z(\boldsymbol{x}) - Z(\boldsymbol{x}+\boldsymbol{h})\right]^2\right\} = \frac{1}{2}Var\left[Z(\boldsymbol{x}) - Z(\boldsymbol{x}+\boldsymbol{h})\right], \tag{25}$$

cuja função, que é a média aritmética do quadrado das diferenças do todos os pares de observações que estão separados por um vetor $\boldsymbol{h}$, pode ser estimada por

$$\hat{\gamma}(\boldsymbol{h}) = \frac{1}{2N(\boldsymbol{h})}\sum_{i=1}^{n(\boldsymbol{h})}\left[z(\boldsymbol{x}_i) - z(\boldsymbol{x}_i+\boldsymbol{h})\right]^2. \tag{26}$$

A Figura 14 mostra uma representação do semivariograma experimental utilizado na análise da estrutura da variação espacial de uma variável regionalizada. Segundo Camargo

(1997), o patamar (C) é uma estimativa da variância do conjunto de dados, representando o ponto a partir do qual não existe mais dependência espacial entre as observações, ou seja, termina o efeito da distância sobre a variância entre os pares de observações. O ponto em que a variância se torna constante é definido como o alcance (a), ou seja, o alcance identifica a distância dentro da qual as observações apresentam correlação espacial. Além desse ponto predominam os efeitos de aleatoriedade. Já o efeito pepita (C_0), identificado como sendo a intersecção da curva ajustada do semivariograma no eixo das ordenadas, representa a variação não identificada por estar estruturada a distâncias menores que a distância entre as observações. Trata-se da chamada variação de pequena escala, que também contem os erros de medição. A contribuição (C_1) é a diferença entre o patamar e o efeito pepita, e representa a semivariância espacialmente estruturada.

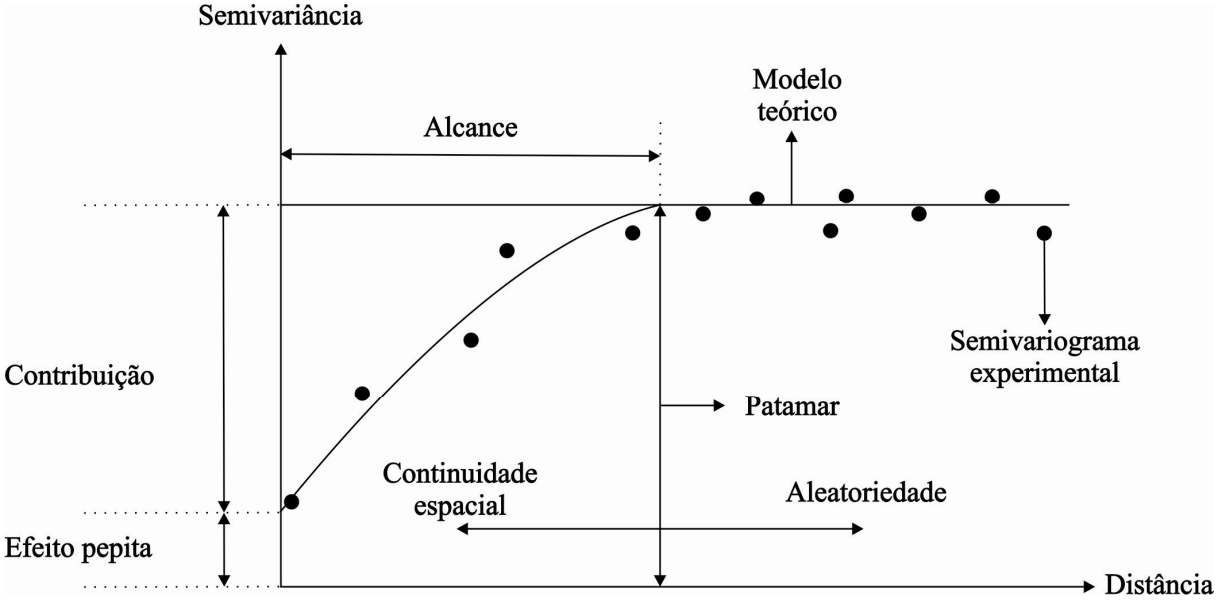


Figura 14 – Representação esquemática de um semivariograma (adaptado de Camargo (1997)).

A análise geoestatística dos res_i foi conduzida no ArcGIS (ESRI, 2009) utilizando o *Geostatistical Analyst*. Para avaliar a proporção da variância explicada pelo modelo ajustado, utilizei o Grau de Dependência Espacial (*GDE*). O *GDE* é dado por

$$GDE = \frac{C_1}{C_0 + C_1}. \quad (27)$$

onde *GDE* é o Grau de Dependência Espacial (%), C_1 é a contribuição, C_0 é o efeito pepita, que somados definem o patamar. A interpretação do *GDE* foi feita conforme (Cambardella et al., 1994): fraca se *GDE* < 25%, moderada se 25% < *GDE* < 75%, e forte se GDE > 75%. Em sendo verificada a existência de dependência espacial, os resíduos foram krigados, gerando mapas que foram utilizados para avaliar espacialmente o desempenho das FPESe.

5.5.3 Validação das FPESe

O segundo objetivo do uso de subconjuntos de observações selecionadas aleatoriamente (*n* = 300; *n* = 150; *n* = 150) para o ajuste dos modelos de regressão foi favorecer a eficiência do procedimento de validação dos modelos de regressão ajustados. Isso porque a validação de modelos de regressão ajustados deve ser realizada sem utilizar observações que tenham sido previamente utilizadas para ajustá-lo (Brus et al., 2011). Caso contrário os resultados serão enviesados. Assim, modelos de regressão ajustados com dados provenientes de procedimentos amostrais intencionais devem ser validados utilizando um conjunto de dados amostrados aleatoriamente (Brus et al., 2011). Contudo, devido a dificuldades de acesso e escassez de força de trabalho e recursos financeiros, não pude realizar tal amostragem aleatória. A alternativa escolhida foi a validação cruzada (*cross-validation*) (GlobalSoilMap.net, 2011), procedimento no qual todas as observações (*n* = 339; *n* = 165; *n* = 174) foram utilizadas. Com isso introduzi no procedimento *n* = 39, *n* = 15 e *n* = 24 observações desconhecidas dos modelos de regressão ajustados. Reconheço tratar-se de conjuntos pequenos de observações, mas foi a melhor alternativa que possuía em mãos. A outra alternativa disponível era a partição dos conjuntos iniciais de dados (*n* = 339; *n* = 165; *n* = 174) em dois: um conjunto de teste e um conjunto de validação (Brus et al., 2011). Contudo, considerei o tamanho do conjunto de dados disponível inadequado para tal procedimento. Rawlins et al. (2009), por exemplo, utilizaram 825 observações para validar seu modelo de regressão linear múltipla. Mas permanece, conforme já dito acima, a necessidade de avaliar a efetividade do procedimento que adotei.

O procedimento da validação cruzada consiste (Varmuza e Filzmoser, 2009) na partição aleatória do conjunto de dados utilizado para o ajuste dos modelos de regressão em *S* segmentos com aproximadamente o mesmo tamanho (número de observações). O valor de *S* pode variar entre 2 e *n*, mas geralmente valores entre 4 e 10 são utilizados. A Figura 15 demonstra o esquema de uma validação cruzada com quatro segmentos. A cada passo do procedimento um segmento é deixado "separado" para ser utilizado como conjunto de validação. Os demais $S - 1$ segmentos são utilizados como conjunto de teste, ou seja, para ajustar os modelos de regressão utilizando as variáveis preditoras definidas acima (CONV, ELEV, IPE e IUT). Depois de ajustados os modelos de regressão utilizando esse conjunto de observações dos $S - 1$ segmentos, faz-se a predição dos valores das observações do conjunto deixado "separado", o conjunto de validação. Esse procedimento é repetido até que cada segmento *S* seja, em algum momento, deixado "separado" para constituir o conjunto de dados de validação enquanto os outros $S - 1$ segmentos são utilizados para ajustar os modelos de regressão. A partir das predições realizadas para cada um dos *S* segmentos são calculados os resíduos conforme a Equação (18).

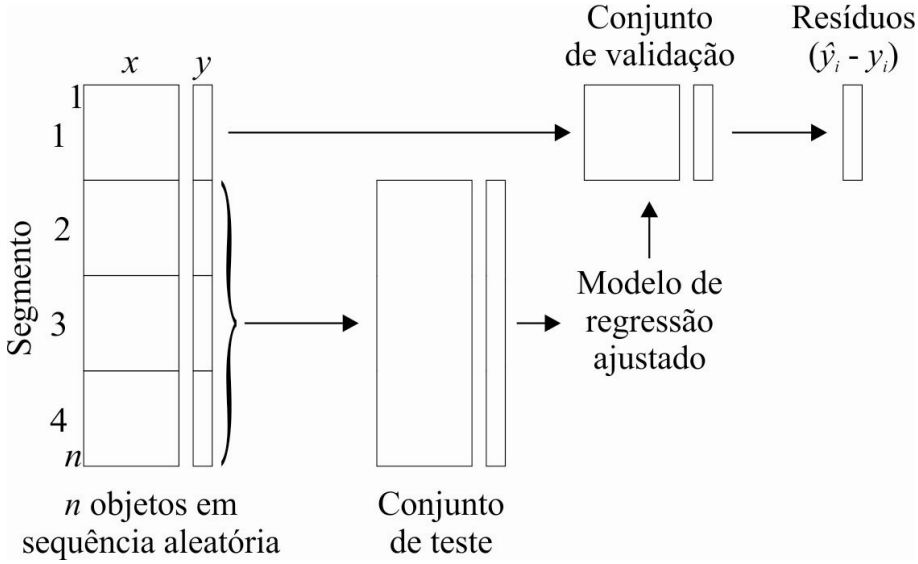


Figura 15 – Representação esquemática do primeiro passo de uma validação cruzada onde o conjunto de dados foi dividido em quatro segmentos (adaptado de Varmuza e Filzmoser (2009)).

A realização de um único procedimento de validação cruzada resulta em *n* predições. Contudo, isso não permite a visualização e avaliação da distribuição do erro. Além disso, as estatísticas de avaliação do desempenho dos modelos de regressão ajustados dependerão, sobretudo, do número de segmentos *S* em que o conjunto de dados foi dividido. Devido a esses problemas (Varmuza e Filzmoser, 2009) recomendam que o procedimento de validação cruzada seja repetido utilizando diferentes partições aleatórias. Isso significa que a cada nova rodada de validação cruzada uma nova partição aleatória do conjunto de dados é realizada (Refaeilzadeh et al., 2009) e o procedimento preditivo descrito acima repetido. Ao final da análise cada observação do conjunto de dados terá a ela associado um número de predições igual ao número de vezes que a validação cruzada foi repetida. Esses resultados podem ser utilizados para calcular a distribuição das estatísticas que descrevem os erros de predição (Rawlins et al., 2009), o que permite uma avaliação mais acurada da qualidade dos modelos de regressão ajustados.

Com base nas recomendações acima descritas, a validação dos modelos de regressão ajustados foi realizada utilizando uma validação cruzada com os conjuntos de dados particionados aleatoriamente em dez segmentos (*10-fold cross validation*), repetida 100 vezes, utilizando o comando `train()` implementado no pacote *caret* (Kuhn et al., 2012). As partições foram selecionadas aleatoriamente, utilizando como semente aleatória o número 123. Mais uma vez, não há qualquer critério estatístico ou matemático que fundamente a utilização do número 123 como semente aleatória. Sua escolha se deu por se tratar de uma sequência numérica de fácil memorização. Em cada rodada de validação cruzada, os valores preditos das log-razões aditivas *ln*(*argila*/*areia*) e *ln*(*silte*/*areia*) foram transformados para a escala original (argila, silte e areia) conforme preconizado por Aitchison (1982) (Equação (7)).

O conjunto de dados da validação cruzada (100 valores preditos para cada observação) dos modelos de regressão ajustados para o conjunto de dados com $n = 339$ observações foi particionado em dois. Como critério utilizei o valor de elevação associado a cada observação. Assim, os dados associados às observações localizadas em elevações > 300 m foram agrupados em um primeiro subconjunto, enquanto os dados associados às observações localizadas em elevações ≥ 300 m foram agrupados em um segundo subconjunto. Essa divisão teve como objetivo avaliar as predições feitas pelos modelos de regressão ajustados com $n = 300$ observações (representativas de toda a área de estudo) nos domínios inferior e superior da área de estudo. Além disso, objetivei comparar essas predições com aquelas feitas pelos modelos de regressão ajustados especificamente para cada domínio fisiográfico.

De posse dos cinco conjuntos de dados, calculei quatro estatísticas para cada fração de tamanho de partícula: o erro médio, a raiz quadrada do erro quadrático médio, a raiz quadrada do erro quadrático médio normalizada e o coeficiente de determinação ajustado.

O erro médio é dado por

$$EM = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right), \tag{28}$$

onde *EM* é o erro médio, $\hat{y}_i$ é o valor predito na *i*-ésima observação e y_i é o valor medido na *i*-ésima observação, com *i* = 1, 2, ..., *n* (McBratney et al., 2011). Valores positivos indicam sobrestimativa e valores negativos indicam subestimativa (McBratney et al., 2011).

O erro quadrático médio é dado por

$$EQM = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2, \tag{29}$$

onde *EQM* é o erro quadrático médio, $\hat{y}_i$ é o valor predito na *i*-ésima observação e y_i é o valor medido na *i*-ésima observação, com *i* = 1, 2, ..., *n* (Wikipedia, 2012b). Para um estimador não enviesado, o *EQM* é a variância, onde valores mais próximos de zero indicam que a acurácia das predições é elevada (Wikipedia, 2012b).

A raiz quadrado do erro quadrático médio é dada por

$$RMSE = \sqrt{EQM}, \tag{30}$$

onde *RMSE* é a raiz quadrada do erro quadrático médio e *EQM* é o erro quadrático médio (Wikipedia, 2012). Para um estimador não enviesado, a *RMSE* é o desvio padrão, onde valores mais próximos de zero indicam que a acurácia das predições é elevada (Wikipedia, 2012c).

A raiz quadrada do erro quadrático médio normalizada é dada por

$$RMSE_{norm} = \frac{RMSE}{y_{\max} - y_{\min}}, \tag{31}$$

onde $RMSE_{norm}$ é a raiz quadrada do erro quadrático médio normalizada, $y_{\max}$ é o maior valor medido e $y_{\min}$ é o menor valor medido (Wikipedia, 2012c). Ao normalizar o *RMSE* é possível comparar a qualidade das predições das diferentes frações de tamanho de partícula do solo. Quanto menor for o $RMSE_{norm}$, menor será a variância contida nos resíduos (Wikipedia, 2012c).

O coeficiente de determinação ajustado é dado por

$$R_{aj}^2 = 1 - \frac{n-1}{n-(k+1)}\left(1 - R^2\right), \tag{32}$$

onde R_{aj}^2 é o coeficiente de determinação ajustado, n é o número de observações utilizadas para ajustar o modelo de regressão, k é o número de variáveis preditoras utilizadas no ajuste do modelo de regressão e R^2 é o coeficiente de determinação (Wikipedia, 2012a). A vantagem do coeficiente de determinação ajustado R_{aj}^2 é que ele penaliza a inclusão de variáveis pouco explicativas no modelo. Isso é importante porque o coeficiente de determinação R^2 é influenciado pelo número de variáveis preditoras relativo ao número de observações utilizado para ajustar o modelo (Hair et al., 2010). Além disso, o coeficiente de determinação ajustado R_{aj}^2 permite que comparemos modelos de regressão ajustados utilizando diferentes números de variáveis preditoras e observações (Hair et al., 2010), algo que o coeficiente de determinação R^2 não permite.

O coeficiente de determinação é dado por

$$R^2 = \frac{SQ_{\text{explicada}}}{SQ_{total}}, \tag{33}$$

onde R^2 é o coeficiente de determinação, SQ_{total} é a soma de quadrados total e $SQ_{explicada}$ é a soma de quadrados da explicada pelo modelo de regressão ajustado (igual a $SQ_{regressão}$) (Wikipedia, 2012a).

A soma de quadrados total é dada por

$$SQ_{total} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{34}$$

onde SQ_{total} é a soma de quadrados total, y_i é o valor medido na i-ésima observação e $\bar{y}$ é a média aritmética dos valores medidos, com $i = 1, 2, ..., n$ (Wikipedia, 2012a).

A soma de quadrados da regressão é dada por

$$SQ_{\text{explicada}} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 , \tag{35}$$

onde $SQ_{explicada}$ é a soma de quadrados explicada pelo modelo de regressão ajustado, $\hat{y}_i$ é o valor predito na i-ésima observação e $\bar{y}$ é a média aritmética dos valores medidos, com $i = 1, 2, ..., n$ (Wikipedia, 2012a).

Para cada uma das quatro estatísticas (*EM*, *RMSE*, $RMSE_{norm}$ e R^2_{aj}) foram obtidos os percentis 2,5 e 97,5 e a mediana (Rawlins et al., 2009) utilizando os comandos `quantile()` e `median()` implementados no pacote *stats* (R Development Core Team, 2011).

6 RESULTADOS

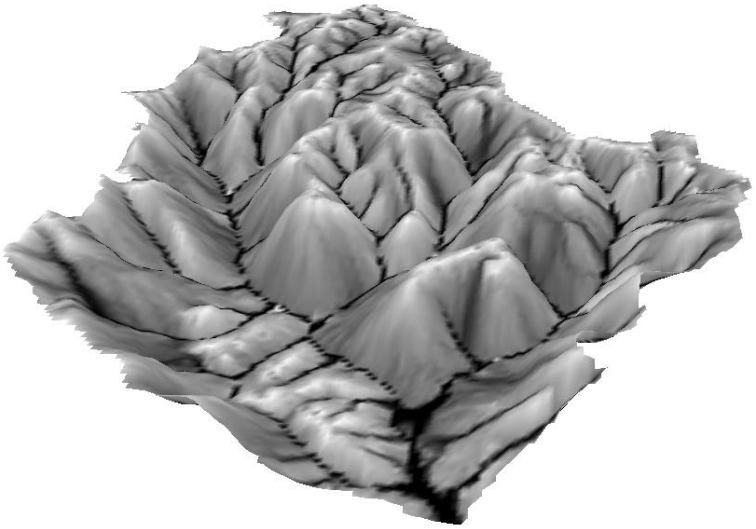
6.1 Variáveis preditoras

A Tabela 1 apresenta as quinze variáveis preditoras utilizadas, a simbologia adotada nesse estudo, o seu coeficiente de assimetria e as transformações realizadas, enquanto a Figura 16 apresenta a sua distribuição espacial na área de estudo.

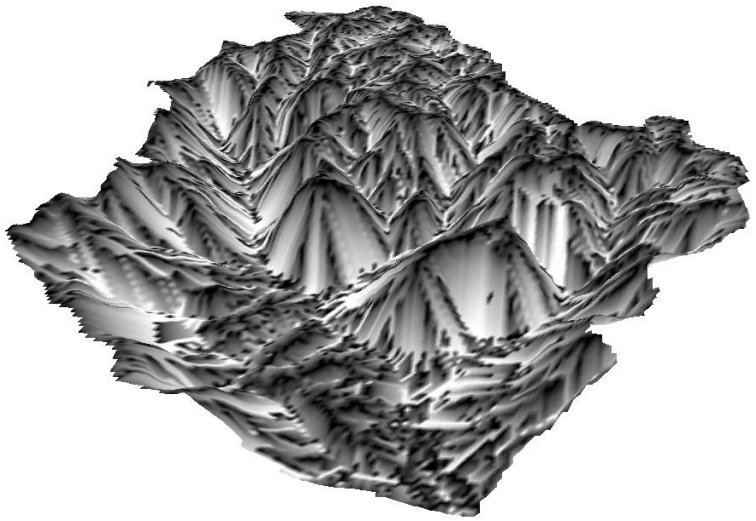
Tabela 1 – As quinze variáveis preditoras, a simbologia adotada, seus coeficientes de assimetria[1] e as transformações realizadas para obter uma distribuição próxima da normal.

Variáveis preditoras e simbologia utilizada	Coeficiente de assimetria	Escala original	Transformação	
			Raiz quadrada	Logaritmo natural
Atributos primários				
Área de Contribuição (AC)	12,68			x
Comprimento do Declive (CD)	2,93		x	
Curvatura (CURV)	-0,31	x		
Curvatura de Perfil (PERF)	0,40	x		
Curvatura Planar[2] (PLAN)	-1,60	x		
Declividade (DECL)	0,64		x	
Declividade Média da Área de Contribuição (DMAC)	0,35	x		
Elevação (ELEV)	0,03	x		
Elevação Acima da Rede de Drenagem (EARD)	1,79		x	
Northerness (NORT)	0,03	x		
Atributos secundários				
Fator LS (LS)	3,35		x	
Índice de Convergência (CONV)	-0,13	x		
Índice de Potência de Escoamento[3] (IPE)	10,52			x
Índice de Rugosidade da Superfície (IRS)	0,78		x	
Índice de Umidade Topográfica (IUT)	2,58			x

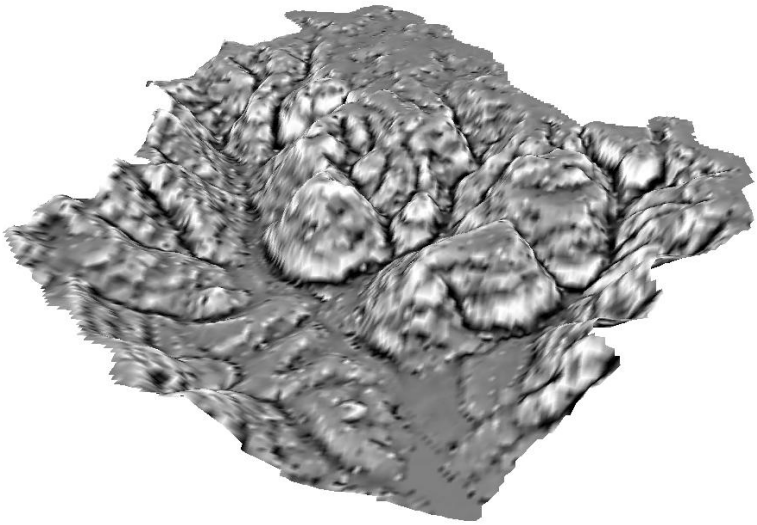
[1] Coeficiente de assimetria calculado com base nos 339 pontos amostrados antes da transformação dos dados.
[2] A presença de valores negativos impede a transformação usando as funções raiz quadrada e logaritmo natural.
[3] Devido à ocorrência de valores de IPE iguais a zero (0,0) um valor igual a um (1,0) foi somado em cada observação para permitir a transformação dos dados para a escala logarítmica.



Estatísticas descritivas de AC	
Média aritmética	7,46
Desvio padrão	1,45
Valor mínimo	4,60
Valor máximo	16,64

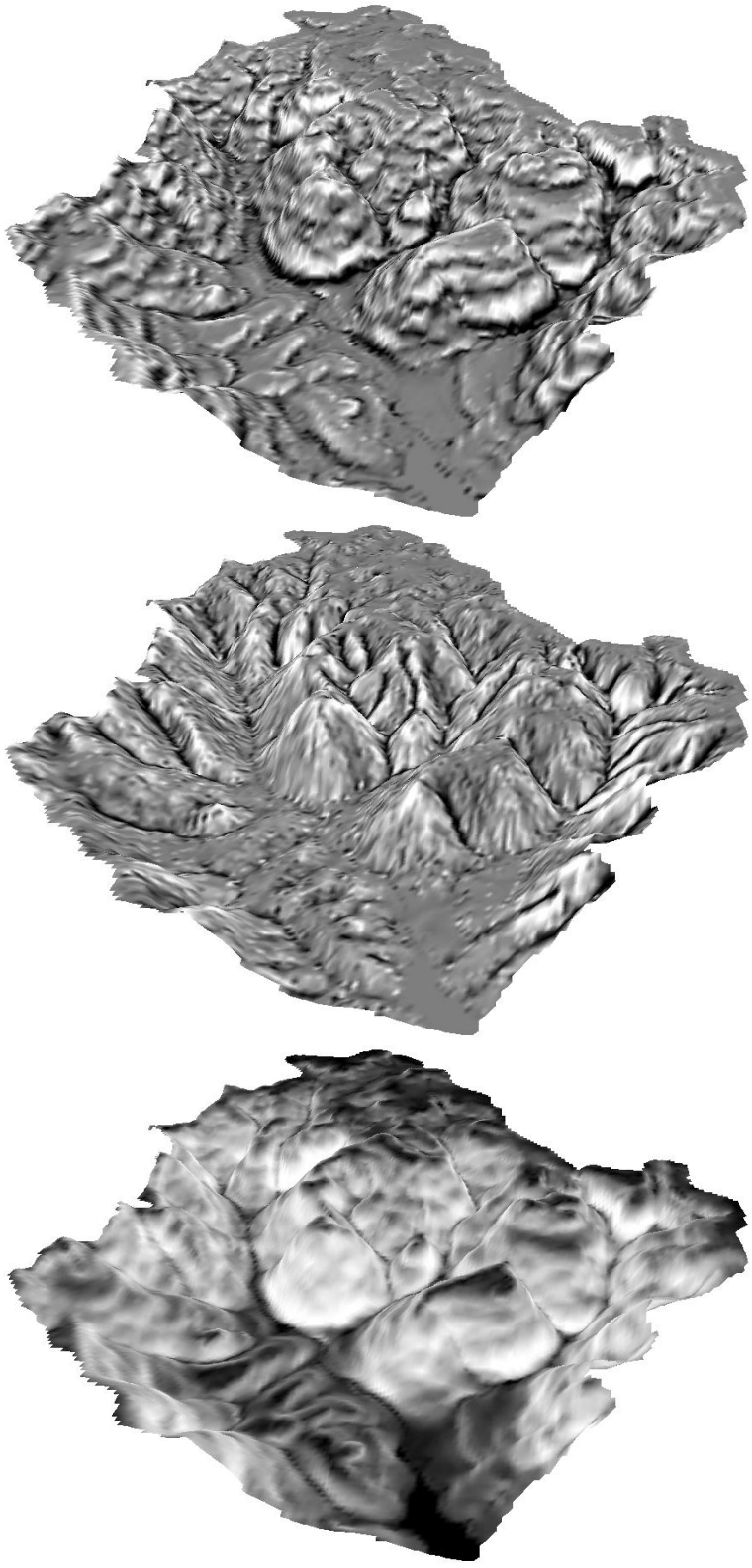


Estatísticas descritivas de CD	
Média aritmética	9,55
Desvio padrão	5,53
Valor mínimo	0,00
Valor máximo	42,40



Estatísticas descritivas de CURV	
Média aritmética	-0,00
Desvio padrão	0,01
Valor mínimo	-0,05
Valor máximo	0,04

Figura 16 – Perspectiva em 3D da distribuição espacial das variáveis preditoras[1] (atributos de terreno) na área de estudo e suas estatísticas descritivas (314.226 células de 100 m²). Continua...

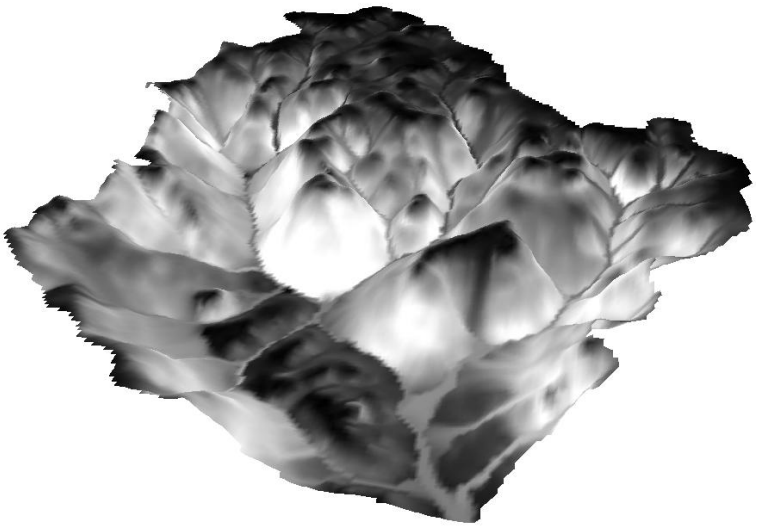


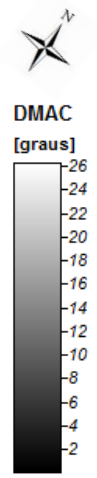
Estatísticas descritivas de PERF	
Média aritmética	-0,00
Desvio padrão	0,00
Valor mínimo	-0,04
Valor máximo	0,03

Estatísticas descritivas de PLAN	
Média aritmética	0,00
Desvio padrão	0,00
Valor mínimo	-0,04
Valor máximo	0,03

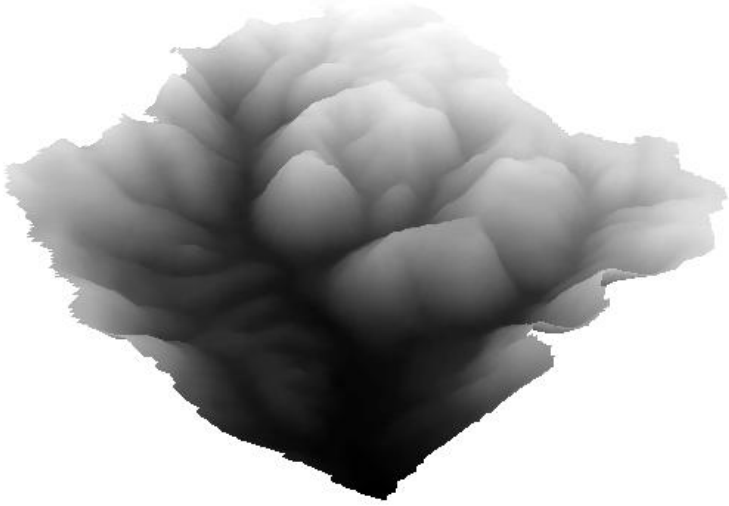
Estatísticas descritivas de DECL	
Média aritmética	3,26
Desvio padrão	1,39
Valor mínimo	0,00
Valor máximo	6,90

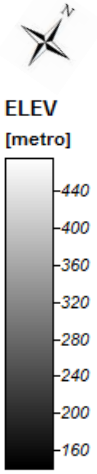
Figura 16 – Continuação.



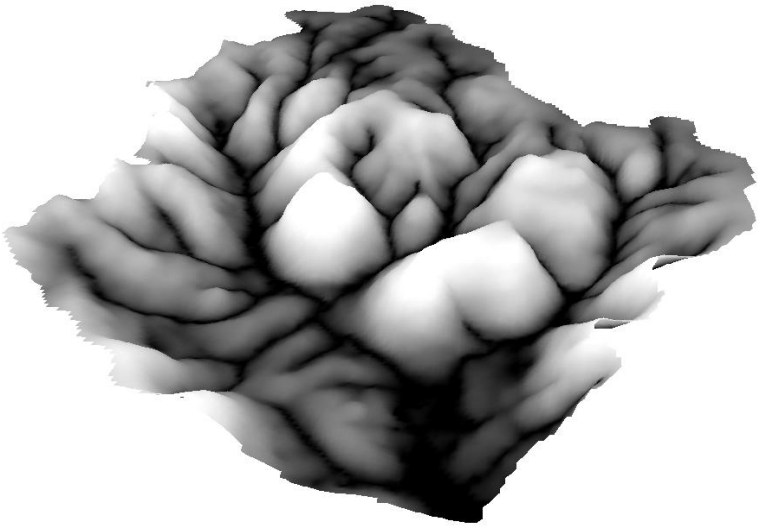


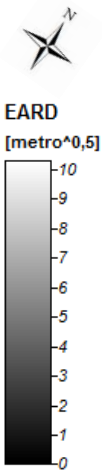
Estatísticas descritivas de DMAC	
Média aritmética	11,48
Desvio padrão	7,36
Valor mínimo	0,01
Valor máximo	35,86





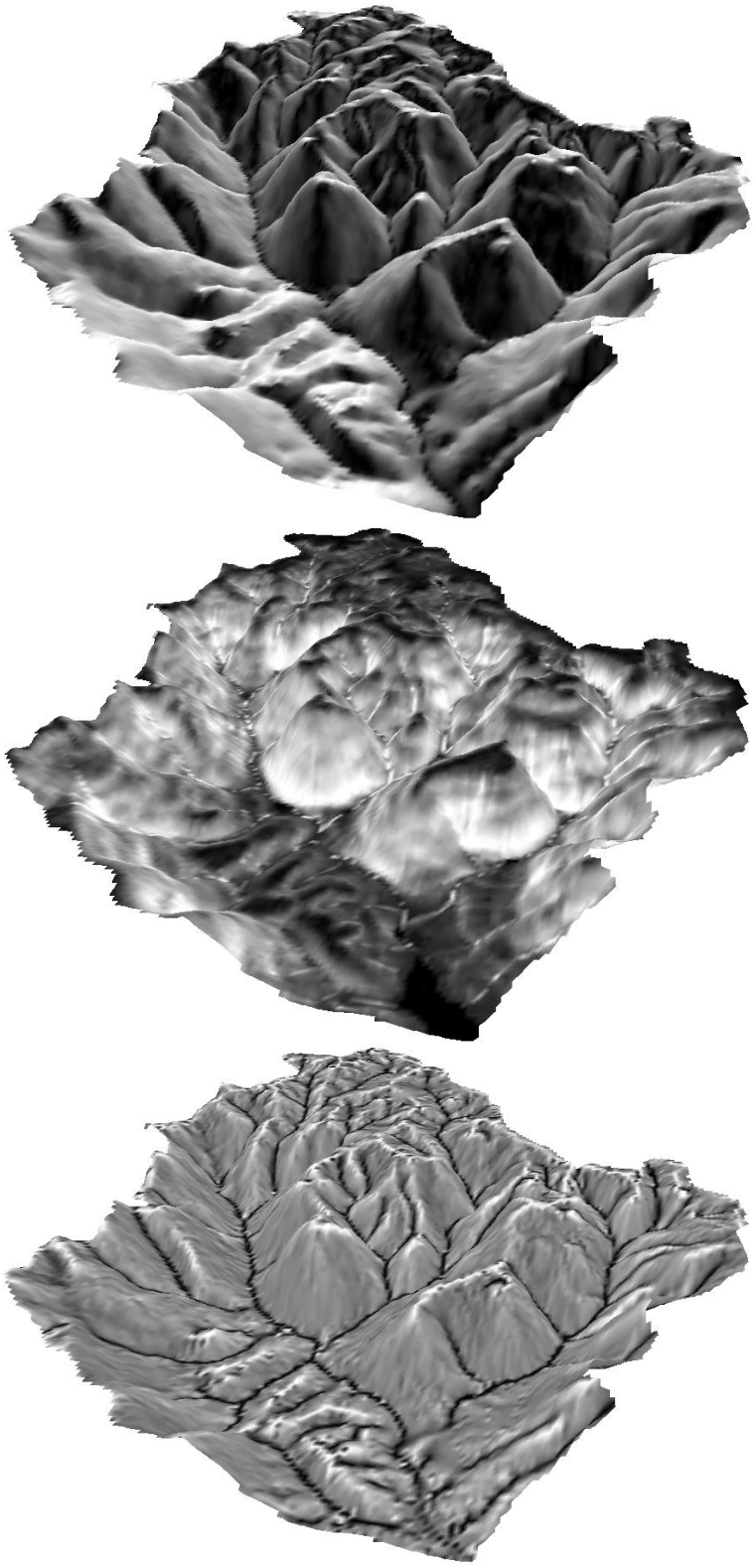
Estatísticas descritivas de ELEV	
Média aritmética	331,73
Desvio padrão	96,91
Valor mínimo	139,18
Valor máximo	475,18





Estatísticas descritivas de EARD	
Média aritmética	4,59
Desvio padrão	2,84
Valor mínimo	0,00
Valor máximo	13,40

Figura 16 – Continuação.

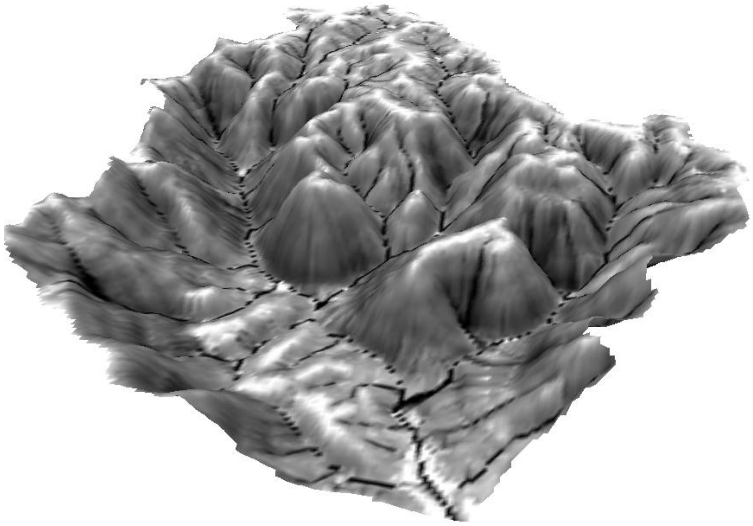


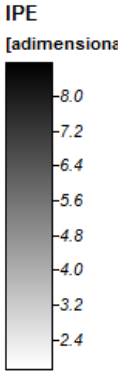
Estatísticas descritivas de NORT	
Média aritmética	78,57
Desvio padrão	47,19
Valor mínimo	0,00
Valor máximo	180,00

Estatísticas descritivas de LS	
Média aritmética	4,49
Desvio padrão	2,57
Valor mínimo	0,00
Valor máximo	30,53

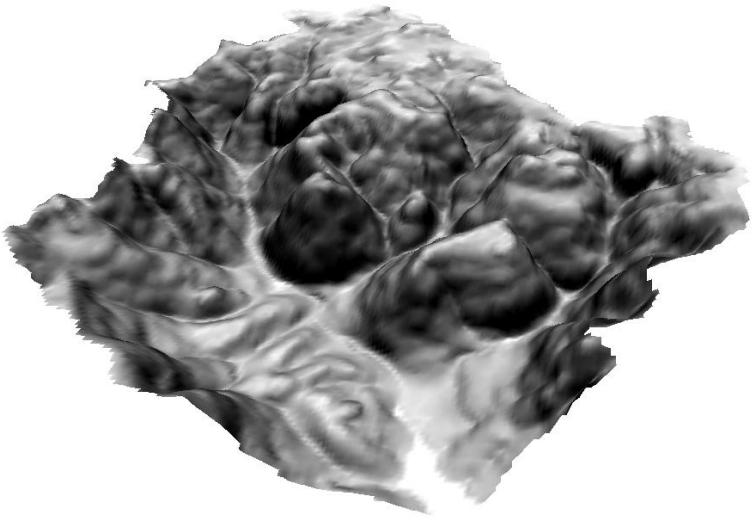
Estatísticas descritivas de CONV	
Média aritmética	0,06
Desvio padrão	9,74
Valor mínimo	-82,12
Valor máximo	90,89

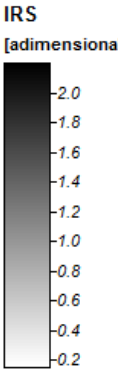
Figura 16 – Continuação.



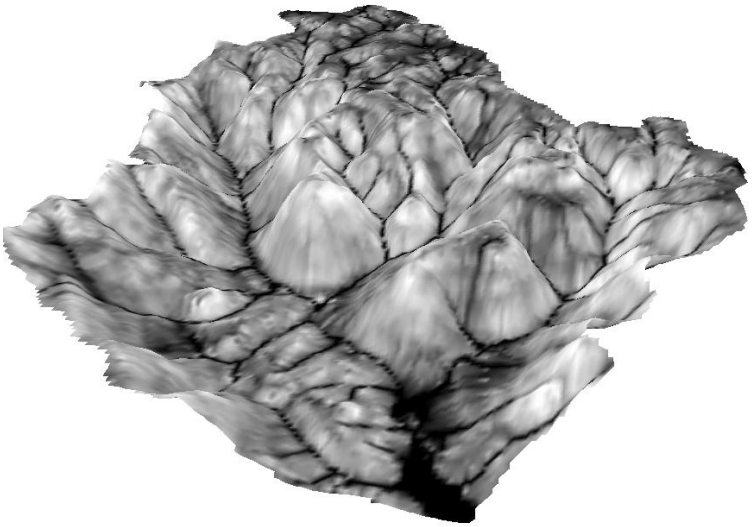


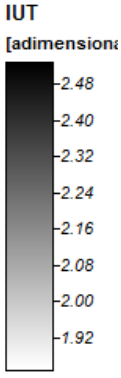
Estatísticas descritivas de IPE	
Média aritmética	5,23
Desvio padrão	1,76
Valor mínimo	-2,30
Valor máximo	13,57





Estatísticas descritivas de IRS	
Média aritmética	1,17
Desvio padrão	0,51
Valor mínimo	0,03
Valor máximo	2,78





Estatísticas descritivas de IUT	
Média aritmética	2,19
Desvio padrão	0,17
Valor mínimo	1,77
Valor máximo	3,16

Figura 16 – Continuação.

[1] AC – logaritmo natural da área de contribuição, CD – raiz quadrada do comprimento do declive, CURV – curvatura, PERF – curvatura de perfil, PLAN – curvatura planar, DECL – raiz quadrada da declividade, DMAC – declividade média da área de contribuição, ELEV – elevação, EARD – raiz quadrada da elevação acima da rede de drenagem, NORT – northerness, LS – raiz quadrada do fator LS, CONV – índice de convergência, IPE – logaritmo natural do índice de poder de escoamento, IRS – raiz quadrada do índice de rugosidade da superfície, IUT – logaritmo natural do índice de umidade topográfica.

Note que 50% das variáveis preditoras precisou ser transformada para alcançar uma distribuição mais próxima da normal. Cinco delas foi transformada para sua raiz quadrada e quatro delas para a escala logarítmica natural. A variável PLAN também deveria ter sido transformada, mas devido à ocorrência de valores negativos as duas opções de transformação utilizadas não são adequadas. Problema similar ocorreu com IPE, que apresentou valores iguais a 0,0. Nesse caso acresci um valor igual a 1,0 para permitir a transformação dos dados.

A Figura 17 apresenta os histogramas de frequência das variáveis preditoras (construídos com base nos 339 pontos amostrados dos planos de informação) na escala em que foram utilizadas nas análises subsequentes (após a transformação quando necessário) e as suas estatísticas descritivas aparecem na Tabela 2. Esses dados mostram que a informação amostrada para os 339 pontos de coleta cobre a maior parte do intervalo de variação dos atributos de terreno na área de estudo.

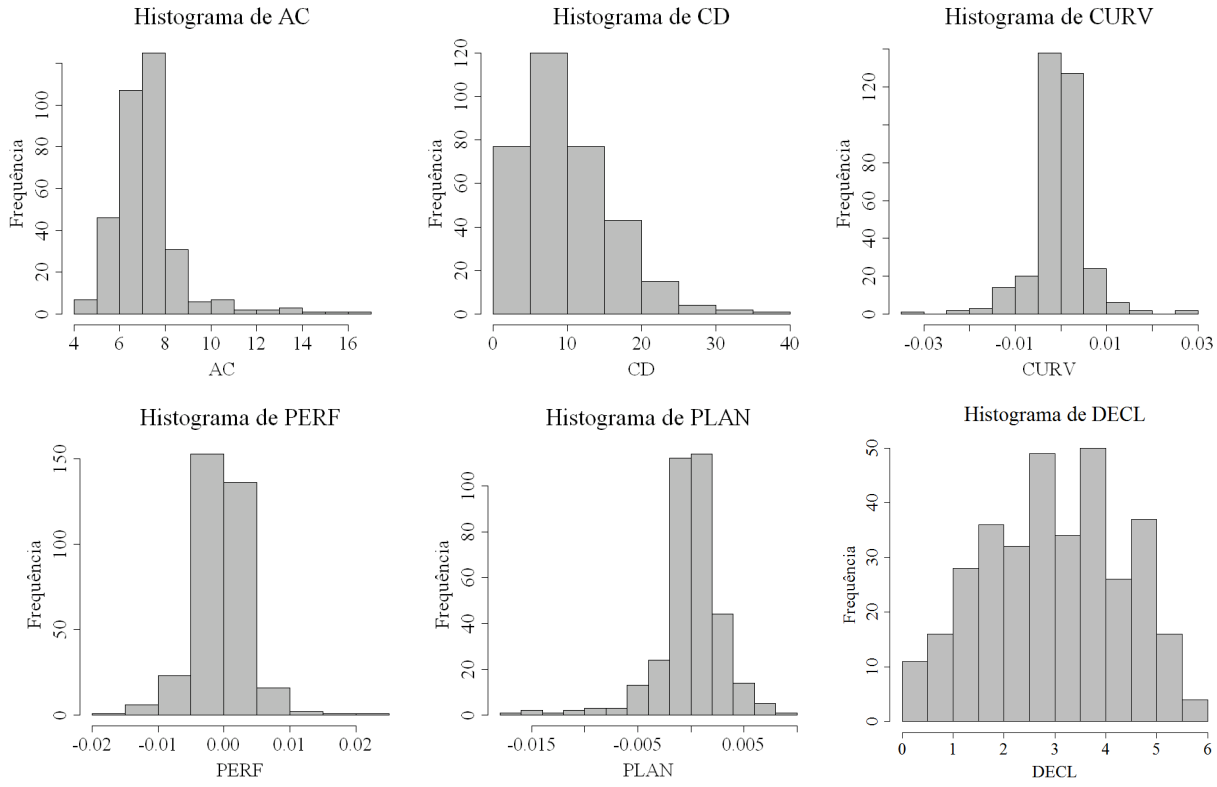


Figura 17 – Histogramas de frequência das variáveis preditoras[1] utilizadas na construção das funções de predição espacial da distribuição do tamanho de partículas do solo. Continua...

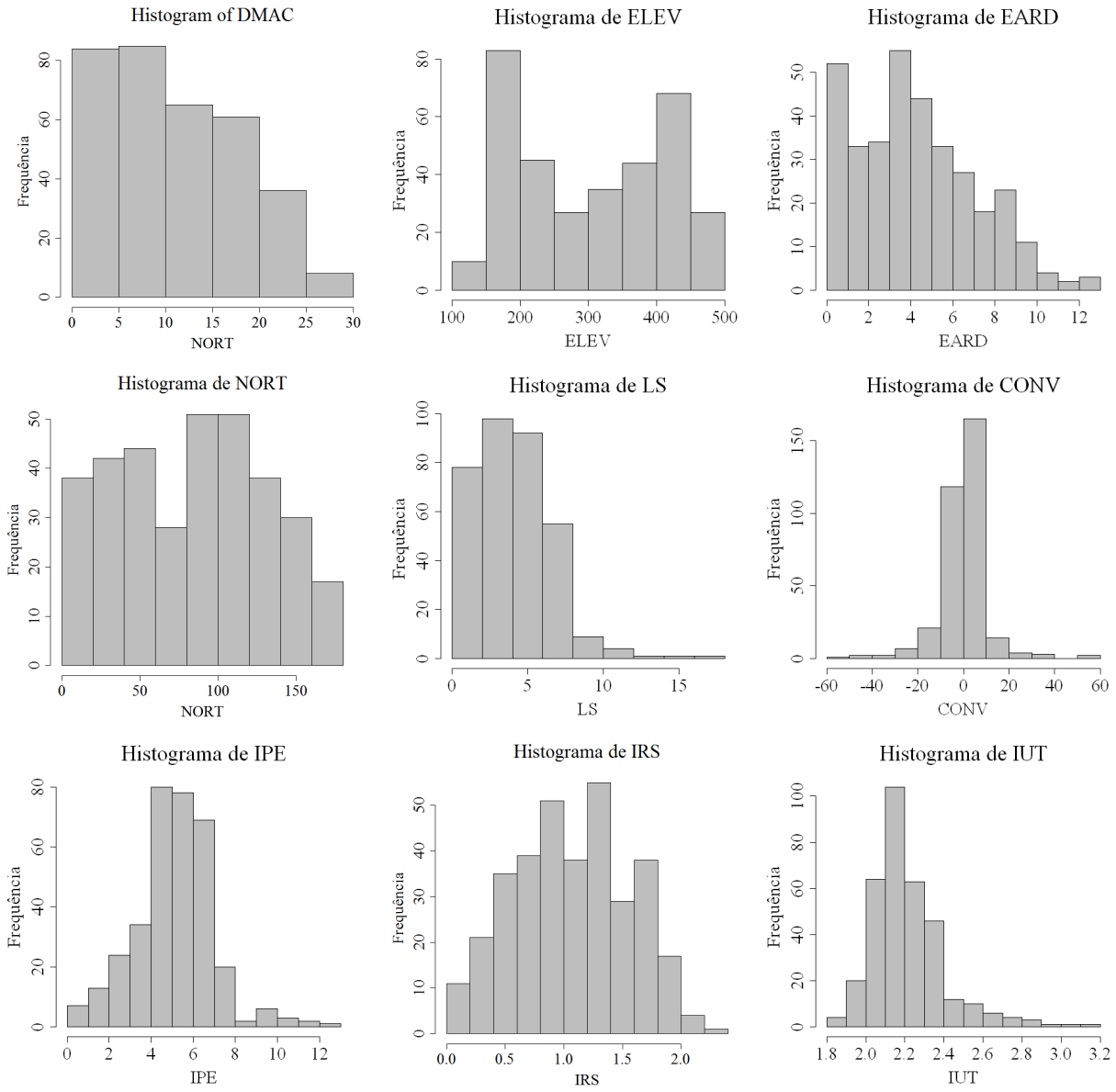


Figura 17. Continuação...

[1] AC – logaritmo natural da área de contribuição, CD – raiz quadrada do comprimento do declive, CURV – curvatura, PERF – curvatura de perfil, PLAN – curvatura planar, DECL – raiz quadrada da declividade, DMAC – declividade média da área de contribuição, ELEV – elevação, EARD – raiz quadrada da elevação acima da rede de drenagem, NORT – northerness, LS – raiz quadrada do fator LS, CONV – índice de convergência, IPE – logaritmo natural do índice de poder de escoamento, IRS – raiz quadrada do índice de rugosidade da superfície, IUT – logaritmo natural do índice de umidade topográfica.

Entretanto, como a amostragem dos solos na área de estudo foi baseada no conhecimento tácito dos pedólogos (seção 5.2), algumas feições do terreno acabaram sendo subamostradas. A comparação dos dados das Figura 16 e Figura 17 e da Tabela 2 evidencia esse problema. Em resumo, o conjunto de dados das variáveis preditoras que melhor captura a variabilidade dos atributos de terreno na área de estudo são AC, ELEV, EARD, IUT e NORT.

Enquanto isso, os locais de maior comprimento de declive (CD, LS), de maior concavidade (CURV, PERF, PLAN), maior declividade (DECL, DMAC, LS) e maior rugosidade (IRS) foram subamostrados (note que essas áreas são aquelas de mais difícil acesso). Locais de máxima e mínima CONV e IPE também foram subamostrados.

Tabela 2 – Estatísticas descritivas das quinze variáveis preditoras utilizadas para construção das funções de predição espacial da distribuição do tamanho de partículas do solo nos 339 pontos amostrados na área de estudo.

Variáveis preditoras	Média aritmética	Desvio padrão	Valor mínimo	Valor máximo
		Atributos primários		
AC	7,22	1,54	4,61	16,32
CD	9,79	6,56	0,00	37,44
CURV	0,00	0,01	-0,03	0,03
PERF	0,00	0,00	-0,02	0,02
PLAN	0,00	0,00	-0,02	0,01
DECL	2,99	1,35	0,08	5,93
DMAC	10,92	7,09	0,06	29,96
ELEV	304,00	107,38	148,11	474,86
EARD	4,24	2,90	0,00	13,00
NORT	83,53	47,83	0,56	174,70
		Atributos secundários		
LS	4,14	2,55	0,03	16,03
CONV	0,11	10,64	-52,38	56,96
IPE	5,13	1,87	0,10	12,14
IRS	1,07	0,49	0,03	2,23
IUT	2,22	0,19	1,86	3,13

[1] AC – logaritmo natural da área de contribuição, CD – raiz quadrada do comprimento do declive, CURV – curvatura, PERF – curvatura de perfil, PLAN – curvatura planar, DECL – raiz quadrada da declividade, DMAC – declividade média da área de contribuição, ELEV – elevação, EARD – raiz quadrada da elevação acima da rede de drenagem, NORT – northerness, LS – raiz quadrada do fator LS, CONV – índice de convergência, IPE – logaritmo natural do índice de potência de escoamento, IRS – raiz quadrada do índice de rugosidade da superfície, IUT – logaritmo natural do índice de umidade topográfica.

A Tabela 3 apresenta os coeficientes de correlação linear de Pearson entre as quinze variáveis preditoras. Os coeficientes de correlação linear $\geq 0,80$ estão destacados em negrito. Note que quatro coeficientes de correlação são $\geq 0,80$. São aqueles entre as variáveis PERF e CURV ($r = 0,88$), LS e DECL ($r = 0,85$), IRS e DECL ($r = 1,00$), IPE e LS ($r = 0,84$) e IRS e

LS (r = 0,85). Note que PLAN e CURV possuem uma correlação bastante próxima do limite estabelecido (r = 0,79). Como decorrência desse resultado e com base no critério exposto acima (seção 5.4.3), realizei as seguintes alterações no conjunto de variáveis preditoras:

Tabela 3 - Coeficiente de correlação linear de Pearson entre as quinze variáveis preditoras* (em negrito as correlações superiores a 0,80 → valor adotado como limite máximo de colinearidade aceita entre as variáveis preditoras). Continua...

	AC	CD	CURV	PERF	PLAN	DECL	DMAC	ELEV	EARD
AC	1,00								
CD	0,58	1,00							
CURV	-0,43	-0,35	1,00						
PERF	-0,27	-0,27	**0,88**	1,00					
PLAN	-0,49	-0,33	0,79	0,41	1,00				
DECL	-0,02	0,23	-0,04	-0,06	0,00	1,00			
DMAC	0,28	0,41	-0,16	-0,24	0,01	0,71	1,00		
ELEV	-0,32	-0,26	0,15	0,18	0,06	-0,04	-0,40	1,00	
EARD	-0,45	-0,28	0,43	0,39	0,32	0,37	0,05	0,47	1,00
NORT	-0,06	0,03	-0,12	-0,10	-0,12	0,01	-0,02	-0,12	0,01
LS	0,41	0,43	-0,31	-0,21	-0,33	**0,85**	0,70	-0,14	0,09
CONV	-0,71	-0,42	0,48	0,25	0,60	-0,03	-0,12	0,19	0,34
IPE	0,77	0,62	-0,38	-0,25	-0,40	0,59	0,61	-0,25	-0,13
IRS	-0,01	0,23	-0,05	-0,07	-0,01	**1,00**	0,72	-0,04	0,37
IUT	0,74	0,30	-0,36	-0,21	-0,42	-0,64	-0,20	-0,24	-0,60

a) Eliminei PLAN e PERF em favor de CURV, uma vez que CURV agrega as informações contidas em ambas PLAN e PERF (ver seção 5.4.1);
b) Eliminei IRS em favor de DECL e LS, uma vez que as duas últimas variáveis preditoras devem estar mais relacionadas às bases físicas dos processos que possuem influência sobre a distribuição do tamanho de partículas do solo. Além disso, DECL está implícito em IRS (ver seção 5.4.2). A manutenção de ambos na análise viria a constituir uma redundância teórica;
c) Eliminei LS em favor de IPE e DECL. A manutenção de IPE em detrimento de LS se deve ao fato de que na área de estudo devem predominar os processos erosivos em canal, os quais são relacionados ao IPE. O LS está mais relacionado aos processos erosivos em sulco e entre - sulco, pouco atuantes na área de estudo.

Tabela 3 – Continuação.

	NORT	LS	CONV	IPE	IRS	IUT
AC						
CD						
CURV						
PERF						
PLAN						
DECL						
DMAC						
ELEV						
EARD						
NORT	1,00					
LS	0,01	1,00				
CONV	-0,01	-0,29	1,00			
IPE	-0,02	**0,84**	-0,58	1,00		
IRS	0,01	**0,85**	-0,04	0,59	1,00	
IUT	-0,05	-0,20	-0,49	0,16	-0,63	1,00

[1] AC – logaritmo natural da área de contribuição, CD – raiz quadrada do comprimento do declive, CURV – curvatura, PERF – curvatura de perfil, PLAN – curvatura planar, DECL – raiz quadrada da declividade, DMAC – declividade média da área de contribuição, ELEV – elevação, EARD – raiz quadrada da elevação acima da rede de drenagem, NORT – northerness, LS – raiz quadrada do fator LS, CONV – índice de convergência, IPE – logaritmo natural do índice de poder de escoamento, IRS – raiz quadrada do índice de rugosidade da superfície, IUT – logaritmo natural do índice de umidade topográfica.

Dando sequência às análises que realizei para detectar a ocorrência de multicolinearidade, segue na Tabela 4 o resultado do teste de esfericidade de Bartlett. O resultado mostra que as variáveis preditoras são correlacionadas e, portanto, os dados são adequados para a análise de componentes principais.

Tabela 4 – Estatísticas do teste de esfericidade de Bartlett utilizado para verificar a adequação do conjunto de dados a análise de componentes principais.

Estatística do teste χ^2	Valor p	Graus de liberdade
3612.048	0,00	55

Os testes de adequação amostral (Tabela 5) também mostram que o conjunto de dados é adequado a análise de componentes principais. A exceção é a variável preditora NORT, cuja estatística do teste MSA é < 0,50. Por esse motivo eliminei a variável preditora NORT do conjunto de dados.

Tabela 5 – Estatísticas dos testes de adequação amostral KMO (Kaiser-Meyer-Olkin) e MSA (Measure of Sample Adequacy) das variáveis preditoras selecionadas para a construção das funções de predição espacial da distribuição do tamanho de partículas do solo (valores ≤ 0,50 estão destacados em negrito).

Estatística do teste de adequação amostral KMO											
								a) 0,63 (grau de variância comum medíocre)			
								b) 0,64 (grau de variância comum medíocre)			
Estatísticas do teste de adequação amostral MSA											
	AC	CD	CURV	DECL	DMAC	ELEV	EARD	NORT	CONV	IPE	IUT
a)	0,56	0,82	0,61	0,71	0,67	0,60	0,77	**0,25**	0,84	0,55	0,53
b)	0,56	0,82	0,61	0,70	0,67	0,64	0,80	-	0,84	0,55	0,53

Notas:
a) Estatísticas dos testes realizados com o conjunto de onze variáveis preditoras: AC – logaritmo natural da área de contribuição, CD – raiz quadrada do comprimento do declive, CURV – curvatura, DECL – raiz quadrada da declividade, DMAC – declividade média da área de contribuição, ELEV – elevação, EARD – raiz quadrada da elevação acima da rede de drenagem, NORT – northerness, CONV – índice de convergência, IPE – logaritmo natural do índice de poder de escoamento, IUT – logaritmo natural do índice de umidade topográfica.
b) Estatísticas dos testes realizados após a eliminação da variável preditora NORT.

A Figura 18 mostra a relação entre os autovalores de cada componente principal extraída utilizando as dez variáveis preditoras selecionadas. Apenas três componentes possuem autovalores superiores a 1, as quais explicam juntas 78% da variância contida nos dados (Tabela 6). Note que apenas duas variáveis preditoras possuem pesos elevados (destacados em negrito na Tabela 7) nessas componentes principais: ELEV e DECL. Isso indica que essas variáveis podem ser de grande importância na construção das funções de predição espacial da distribuição do tamanho de partículas do solo. Contudo, as variáveis preditoras DECL e IUT possuem pesos elevados na nona componente principal (0,72 e 0,52, respectivamente), indicando que ambas estão correlacionadas. O coeficiente de correlação linear entre essas duas variáveis é de r = -0,64 (Tabela 3). De fato, DECL é utilizado no cálculo de IUT, conforme mostra a equação (14). Devido à maior relevância conceitual de IUT, DECL pode ser eliminada da análise sem que haja perda significativa de informação.

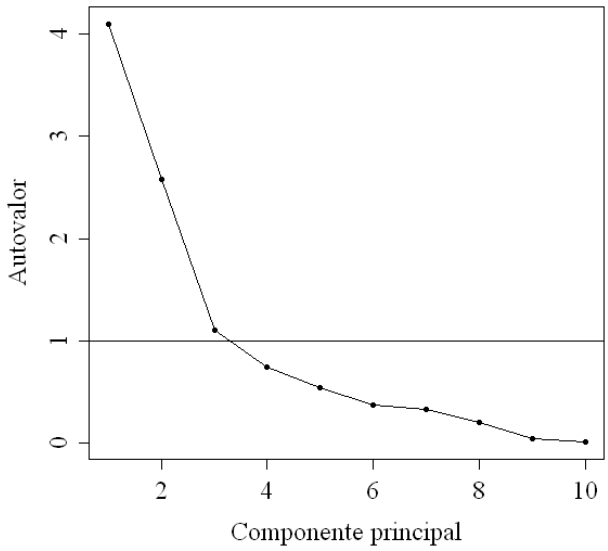


Figura 18 – Autovalor associado a cada componente principal obtida a partir da matriz de correlação das dez variáveis preditoras[1] selecionadas para construção das funções de predição espacial da distribuição do tamanho de partículas do solo (*n* = 339).

[1] Variáveis preditoras: AC – logaritmo natural da área de contribuição, CD – raiz quadrada do comprimento do declive, CURV – curvatura, DECL – raiz quadrada da declividade, DMAC – declividade média da área de contribuição, ELEV – elevação, EARD – raiz quadrada da elevação acima da rede de drenagem, CONV – índice de convergência, IPE – logaritmo natural do índice de poder de escoamento, IUT – logaritmo natural do índice de umidade topográfica.

Tabela 6 – Autovalores e proporção da variância explicada por cada componente principal (*n* = 339).

Componente principal	Autovalor	Proporção da variância explicada	Proporção cumulativa da variância explicada
1	4.104	0.410	0.410
2	2.577	0.258	0.668
3	1.105	0.110	0.779
4	0.739	0.074	0.853
5	0.540	0.054	0.907
6	0.371	0.037	0.944
7	0.321	0.032	0.976
8	0.202	0.020	0.996
9	0.035	0.003	0.999
10	0.006	0.001	1.000

As variáveis preditoras AC e IPE também possuem pesos elevados em uma mesma componente principal, a décima (0,72 e -0,51, respectivamente), mais uma vez indicando a

existência de correlação entre as variáveis. O coeficiente de correlação linear entre essas duas variáveis é de r = 0,77 Tabela 3. De fato o IPE é função da área de contribuição específica (AC_s), conforme mostra a equação (12), que é função de AC (Equação (11)). Devido à maior relevância conceitual de IPE, AC pode ser eliminada da análise sem que haja perda significativa de informação.

Tabela 7 – Pesos das dez variáveis preditoras em cada uma das componentes principais extraídas (os pesos mais elevados de cada variável e todos aqueles ≥ 0,60 estão destacados em negrito) (n = 339).

Variável preditora[1]	Componente principal									
	1	2	3	4	5	6	7	8	9	10
AC	0,45	0,08	-0,16	-0,31	0,12	-0,30	0,12	0,18	-0,01	**0,72**
CD	0,36	-0,13	-0,04	-0,17	**-0,79**	0,38	-0,24	0,00	0,03	0,03
CURV	-0,31	-0,06	0,15	**-0,84**	0,08	0,25	0,28	-0,16	0,04	-0,04
DECL	0,06	**-0,60**	-0,04	0,14	0,10	0,10	0,24	0,13	**0,72**	0,04
DMAC	0,21	-0,47	0,29	-0,02	-0,01	-0,42	-0,09	**-0,66**	-0,17	-0,02
ELEV	-0,24	0,01	**-0,76**	0,06	-0,27	-0,17	0,37	-0,35	-0,02	0,01
EARD	-0,28	-0,33	-0,37	-0,26	0,12	-0,20	**-0,71**	0,21	-0,01	-0,02
CONV	-0,38	-0,06	0,33	-0,02	-0,49	**-0,55**	0,20	0,40	0,00	0,03
IPE	0,39	-0,31	-0,19	-0,14	0,11	-0,07	0,30	0,39	-0,42	**-0,51**
IUT	0,30	0,44	-0,05	-0,24	-0,03	-0,39	-0,12	-0,09	**0,52**	-0,46

[1] Variáveis preditoras: AC – logaritmo natural da área de contribuição, CD – raiz quadrada do comprimento do declive, CURV – curvatura, DECL – raiz quadrada da declividade, DMAC – declividade média da área de contribuição, ELEV – elevação, EARD – raiz quadrada da elevação acima da rede de drenagem, CONV – índice de convergência, IPE – logaritmo natural do índice de poder de escoamento, IUT – logaritmo natural do índice de umidade topográfica.

A Figura 19 mostra a projeção da correlação (Equação (15)) entre as variáveis preditoras e os escores obtidos na análise de componentes principais. Note a formação de cinco agrupamentos distintos. O primeiro deles é formado apenas pela variável ELEV, a qual parece estar agrupada com CURV e CONV no primeiro círculo unitário (projeção da primeira com a segunda componente principal), mas aparece bastante destacada no segundo círculo unitário (projeção da segunda com a terceira componente principal). Já as duas últimas (CONV e CURV) estão agrupadas tanto no primeiro como no segundo círculos unitários, indicando sua estreita relação. O terceiro agrupamento é formado pelas variáveis IUT e EARD, posicionadas exatamente em lados opostos, o que indica sua relação inversa. O quarto

agrupamento é formado pelas variáveis DECL e DMAC, e o quinto pelas variáveis IPE, AC e CD. Como decorrência dos resultados apresentados na Tabela 7 e na Figura 8, as seguintes variáveis podem ser eliminadas das análises subsequentes: AC, CD, CURV, DECL, DMAC e EARD. Portanto, são mantidas para a construção das FPESe as variáveis CONV, ELEV, IPE e IUT.

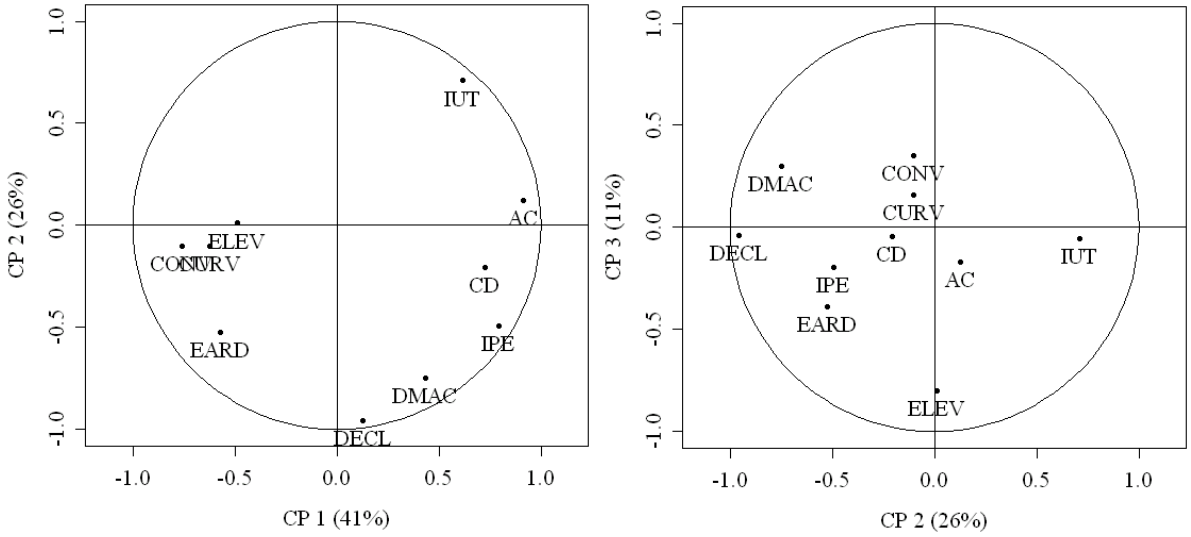


Figura 19 – Projeção da correlação entre as variáveis preditoras[1] e os escores das componentes principais na primeira (CP 1), segunda (CP 2) e terceira (CP 3) dimensões (*n* = 339). Entre parênteses a proporção da variância explicada por cada componente principal.

[1] Variáveis preditoras: AC – logaritmo natural da área de contribuição, CD – raiz quadrada do comprimento do declive, CURV – curvatura, DECL – raiz quadrada da declividade, DMAC – declividade média da área de contribuição, ELEV – elevação, EARD – raiz quadrada da elevação acima da rede de drenagem, CONV – índice de convergência, IPE – logaritmo natural do índice de poder de escoamento, IUT – logaritmo natural do índice de umidade topográfica.

Dentre as quatro variáveis preditoras selecionadas, ELEV é aquela melhor correlacionada às frações de tamanho areia, silte e argila Tabela 8. Todas as demais variáveis possuem coeficientes de correlação inferiores a 0,40. Contudo, a maioria é significativa ao nível de probabilidade de erro de 5%.

Tabela 8 – Coeficientes de correlação linear de Pearson entre as variáveis preditoras[1] selecionadas e as frações da distribuição de tamanho de partículas do solo. Entre parênteses aparecem a estatística do teste t (t) e a sua probabilidade (p) (n = 339).

Variáveis preditoras	Areia	Silte	Argila
ELEV	-0,82 (t = -26,30; p = 0,00)	0,75 (t = 20.82; p = 0,00)	0,74 (t = 20,20; p = 0,00)
CONV	-0,11 (t = -2,03; p = 0,04)	0,07 (t = 1,29; p = 0,20)	0,16 (t = 2,98; p = 0,00)
IPE	0,27 (t = 5,15; p = 0,00	-0,17 (t = -3,17; p = 0,00)	-0,38 (t = -7,54/ p = 0,00)
IUT	0,14 (t = 2,6; p = 0,01)	-0,17 (t = -3,17; p = 0,00)	-0,05 (t = -0,92; p = 0,36)

[1] Variáveis preditoras: ELEV – elevação, CONV – índice de convergência, IPE – logaritmo natural do índice de poder de escoamento, IUT – logaritmo natural do índice de umidade topográfica.

6.2 Funções de predição espacial de solos (FPESe)

6.2.1 FPESe para toda a área de estudo

A FPESe que construí para toda a área de estudo para estimar a log-razão aditiva ln(argila/areia) leva em sua estrutura três variáveis preditoras (ELEV, IPE e CONV) selecionadas através do procedimento stepwise tendo como critério a minimização do *AIC* (Tabela 9). A FPESe construída

$$\ln\left(\frac{argila}{areia}\right) = -2{,}912382 - 0{,}016070 CONV + 0{,}008219 ELEV - 0{,}136263 IPE\,, \qquad (36)$$

onde CONV é o índice de convergência (%), ELEV é a elevação (m) e IPE é o logaritmo natural do índice de potência de escoamento, explica 63% da variância (estimada pela soma de quadrados total) ($R^2 = 0{,}6335$) (Tabela 9 e Tabela 10). A variável que mais contribui para a explicação da variância é ELEV (EV = 59%), seguida de IPE ($EV = 3\%$) e CONV ($EV = 1\%$), conforme calculado através da Equação (21).

Tabela 9 – Coeficientes da função de predição espacial da log-razão aditiva ln(argila/areia) em toda a área de estudo (n = 300).

ln(argila/areia)	Estimativa	Erro padrão	Valor t	Pr(>|t|)
Intercepto	-2,912382	0,202451	-14,386	< 2e-16
CONV[1]	-0,016070	0,004795	-3,351	0,000909
ELEV	0,008219	0,000395	20,807	< 2e-16
IPE	-0,136263	0,027383	-4,976	1,1e-06
Erro quadrático médio: 0,7166 com 296 graus de liberdade				
R^2 múltiplo: 0,6335; R^2 ajustado: 0,6298				
Critério de Informação de Akaike: -195,956				

[1] Variáveis preditoras: ELEV – elevação, CONV – índice de convergência, IPE – logaritmo natural do índice de poder de escoamento.

Note que os coeficientes da FPESe construída indicam que a proporção da fração argila aumenta com o aumento de ELEV (coeficiente com sinal positivo), enquanto a proporção da fração areia aumenta com o aumento de IPE e CONV (coeficiente com sinal negativo) (Tabela 9).

Tabela 10 – Análise de variância da função de predição espacial da log-razão aditiva ln(argila/areia) em toda a área de estudo (n = 300).

Fonte de variação	Graus de liberdade	Soma de quadrados	Quadrado médio	F	Pr(>F)
Regressão[1]	3	262,763	87,588	170,6	< 2,2e-16
CONV	1	5,956	5,956	11,599	0,0007512
ELEV	1	244,090	244,090	475,311	< 2,2e-16
IPE	1	12,717	12,717	24,763	1,102e-06
Resíduos	296	152,007	0,514	-	-
Total	299	414,770	1,387	-	-

[1] Variáveis preditoras: ELEV – elevação, CONV – índice de convergência, IPE – logaritmo natural do índice de poder de escoamento.

A Tabela 11 mostra que essas variáveis preditoras são pouco correlacionadas, uma vez que a variância dos parâmetros do modelo de regressão ajustado pode ser considerada pequena e o FIV é inferior a 2,0. Isso significa que não existe multicolinearidade no conjunto de variáveis preditoras utilizado.

Tabela 11 – Matriz variância-covariância dos parâmetros ajustados da função de predição espacial da log-razão aditiva ln(argila/areia) em toda a área de estudo (n = 300) e fator de inflação da variância (FIV) associado a cada variável preditora.

	Intercepto	CONV	ELEV	IPE	FIV
Intercepto	4,098657e-02	-3,503478e-04	-5,542799e-05	-4,365276e-03	-
CONV[1]	-3,503478e-04	2,299632e-05	-1,167422e-07	7,509373e-05	1,536796
ELEV	-5,542799e-05	-1,167422e-07	1,560092e-07	1,623911e-06	1,058058
IPE	-4,365276e-03	7,509373e-05	1,623911e-06	7,498089e-04	1,566268

[1] Variáveis preditoras: ELEV – elevação, CONV – índice de convergência, IPE – logaritmo natural do índice de poder de escoamento.

A análise dos res_i mostra que os mesmos não estão relacionados aos valores preditos (Figura 20). Contudo, algumas observações podem ser consideradas atípicas e/ou influenciais, sobretudo aquelas com $res_i > |2,0|$, $res_i^* > 1,96$ e $h_{ii} > 3p/n = 3\times4/300 = 0,04$. Mas como os valores de D_i foram todos < 1,0, não removi nenhuma observação do conjunto de dados utilizado no ajuste dos modelos de regressão. Segundo Weisberg (2005), a eliminação de observações com D_i substancialmente inferior a 1,0 não resulta em alteração significativa da estimativa dos parâmetros do modelo ajustado.

Dentre as seis observações que apresentaram os maiores res_i, quatro são provenientes de solos derivados de rochas e materiais sedimentares (observações nº 99, 116, 286 e 328) e as demais de solos derivados de rochas ígneas extrusivas (observações nº 121 e 162) (Tabela 12). Enquanto os valores observados das primeiras são subestimados ($res_i < 0$), os valores observados das últimas são sobreestimados ($res_i > 0$). Esses res_i elevados estão relacionados, principalmente, a localização das observações. As observações provenientes de solos derivados de rochas e materiais sedimentares estão localizadas em ELEV $\geq$ 300 m, onde predominam solos derivados de rochas ígneas extrusivas. Já as observações provenientes de solos derivados de rochas ígneas extrusivas estão localizadas em ELEV < 300 m, onde predominam solos derivados de rochas e materiais sedimentares.

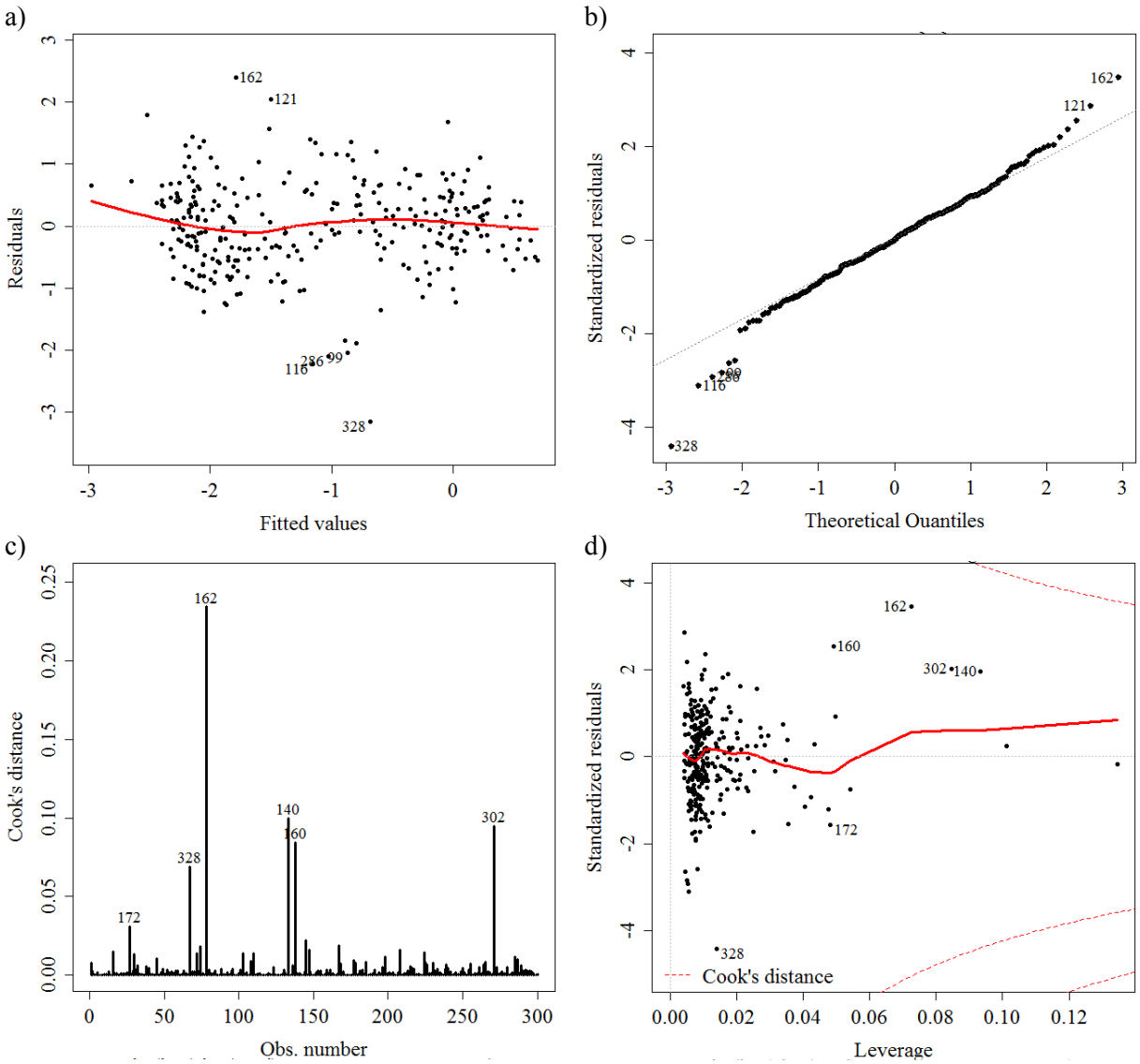


Figura 20 – Análise gráfica da qualidade do ajuste da função de predição espacial da log-razão aditiva ln(argila/areia) em toda a área de estudo (n = 300): a) valores preditos (*fitted values*) e resíduos (*residuals*), b) quantis teóricos (*theoretical quantiles*) e resíduos padronizados (*standardized residuals*), c) distância de Cook (*Cook's distance*) associada a cada observação (*obs. number*), d) alavancagem (*leverage*) e resíduos padronizados (*standardized residuals*).

No que diz respeito às seis observações mais influentes sobre o modelo de regressão ajustado, quatro delas são provenientes de solos derivados de rochas ígneas extrusivas (observações nº 140, 160, 162 e 302). Uma delas apresenta res_i significativo (observação nº 162) e as demais possuem valores de ELEV < 300 m e/ou valores absolutos de CONV e IPE elevados. Das observações provenientes de solos derivados de rochas e materiais sedimentares, uma delas possui res_i significativo (observação nº 328), enquanto a outra é a menos influente das seis (observação nº 172).

Tabela 12 – Identificação das observações atípicas e influenciais na função de predição espacial da log-razão aditiva ln(argila/areia) em toda a área de estudo.

ID	Material de origem	Classe de solo	Areia	Silte	Argila	CONV	ELEV	IPE
99	Sedimentar	RQ	91	4	5	-3,9302	342,751	6,123335
116	Sedimentar	PV	88	9	3	0,518997	319,407	6,30546
121	Ígnea	RL	18	51	31	-1,16179	268,1029	5,963155
140	Ígnea	RL	26	42	32	54,02481	341,8192	1,216597
160	Ígnea	RL	52	23	25	-15,8443	216,2039	12,02097
162	Ígnea	RL	17	52	31	-48,3327	227,5129	11,18767
172	Sedimentar	RF	86	9	5	0	148,1119	0,410245
286	Sedimentar	RQ	91	5	4	1,077544	332,2042	6,06097
302	Ígnea	RL	21	53	26	-52,382	283,0784	10,51904
328	Ígnea	CX	46	53	1	-2,1352	391,5299	7,508506

Nota: ID – identificação da observação, RL – Neossolo Litólico, RQ – Neossolo Quartzarênico, RF – Neossolo Flúvico, PV – Argissolo Vermelho, CX – Cambissolo Háplico, CONV – índice de convergência (%), ELEV – elevação (m), IPE – logaritmo natural do índice de potência de escoamento (adimensional).

A FPESe que construí para toda a área de estudo para estimar a log-razão aditiva ln(silte/areia) leva em sua estrutura as mesmas três variáveis preditoras (ELEV, IPE e CONV) utilizadas para construir a FPESe para estimar a log-razão aditiva ln(argila/areia), também selecionadas através do procedimento *stepwise* tendo como critério a minimização do *AIC* (Tabela 13). A FPESe construída

$$\ln\left(\frac{silte}{areia}\right) = -2{,}8659482 - 0{,}0172176 CONV + 0{,}0091242 ELEV - 0{,}1017627 IPE\,, \qquad (37)$$

onde CONV é o índice de convergência (%), ELEV é a elevação (m) e IPE é o logaritmo natural do índice de potência de escoamento, explica 61% da variância (estimada pela soma de quadrados total) (R^2 = 0,6092) (Tabela 13 e Tabela 14). A variável que mais contribui para a explicação da variância também é ELEV (EV = 59%), seguida de IPE (EV = 1%) e CONV (EV = 1%), conforme calculado através da Equação (21).

Note que os coeficientes da FPESe construída apresentam o mesmo padrão visto acima, ou seja, indicam que a proporção da fração silte aumenta com o aumento de ELEV (coeficiente com sinal positivo), enquanto a proporção da fração areia aumenta com o aumento de IPE e CONV (coeficiente com sinal negativo) (Tabela 13).

Tabela 13 – Coeficientes da função de predição espacial da log-razão aditiva ln(silte/areia) em toda a área de estudo (*n* = 300).

ln(silte/areia)	Estimativa	Erro padrão	Valor t	Pr(>|t|)
Intercepto	-2,8659482	0,2295337	-12,486	< 2e-16
CONV[1]	-0,0172176	0,0054369	-3,167	0,00170
ELEV	0,0091242	0,0004478	20,375	< 2e-16
IPE	-0,1017627	0,0310457	-3,278	0,00117
Erro quadrático médio: 0,8125 com 296 graus de liberdade				
R^2 múltiplo: 0,6092; R^2 ajustado: 0,6053				
Critério de Informação de Akaike: -120,626				

[1] Variáveis preditoras: ELEV – elevação, CONV – índice de convergência, IPE – logaritmo natural do índice de poder de escoamento.

Tabela 14 – Análise de variância da função de predição espacial da log-razão aditiva ln(silte/areia) em toda a área de estudo (*n* = 300).

Fonte de variação	Graus de liberdade	Soma de quadrados	Quadrado médio	F	Pr(>F)
Regressão[1]	3	304,657	101,552	153,8	< 2,2e-16
CONV	1	3,497	3,497	5,2971	0,022056
ELEV	1	294,067	294,067	445,4738	< 2,2e-16
IPE	1	7,093	7,093	10,7442	0,001171
Resíduos	296	195,396	0,660	-	-
Total	299	500,053	1,672	-	-

[1] Variáveis preditoras: ELEV – elevação, CONV – índice de convergência, IPE – logaritmo natural do índice de poder de escoamento.

A Tabela 15 mostra que aqui também essas variáveis preditoras são pouco correlacionadas, uma vez que a variância dos parâmetros do modelo de regressão ajustado pode ser considerada pequena e o FIV é inferior a 2,0. Isso significa, mais uma vez, que não existe multicolinearidade no conjunto de variáveis preditoras utilizado (ELEV, CONV e IPE).

Tabela 15 - Matriz variância-covariância dos parâmetros ajustados da função de predição espacial da log-razão aditiva ln(silte/areia) em toda a área de estudo (n = 300) e fator de inflação da variância (FIV) associado a cada variável preditora.

	Intercepto	CONV	ELEV	IPE	FIV
Intercepto	5,268571e-02	-4,503505e-04	-7,124927e-05	-5,611293e-03	-
CONV[1]	-4,503505e-04	2,956035e-05	-1,500649e-07	9,652836e-05	1,536796
ELEV	-7,124927e-05	-1,500649e-07	2,005402e-07	2,087437e-06	1,058058
IPE	-5,611293e-03	9,652836e-05	2,087437e-06	9,638331e-04	1,566268

[1] Variáveis preditoras: ELEV – elevação, CONV – índice de convergência, IPE – logaritmo natural do índice de poder de escoamento.

Mais uma vez os res_i das predições não estão relacionados aos valores preditos (Figura 21). E apesar de aqui também haverem observações com $res_i > |2{,}0|$, $res_i^* > 1{,}96$ e $h_{ii} > 3p/n = 3\times4/300 = 0{,}04$, não removi nenhuma observação do conjunto de dados para o ajuste dos modelos de regressão. Mais uma vez me pautei pelos valores de D_i, que foram todos < 1,0, ou seja, abaixo do limite proposto por Weisberg (2005).

Dentre as seis observações que apresentaram os maiores res_i, quatro são provenientes de solos derivados de rochas e materiais sedimentares (observações nº 99, 101, 161 e 286) e as demais de solos derivados de rochas ígneas extrusivas (observações nº 135 e 162) (Tabela 16). Enquanto os valores observados das primeiras são subestimados ($res_i < 0$), os valores observados das últimas são sobreestimados ($res_i > 0$). Note que três delas apresentaram res_i significativos na FPESe anterior (observações nº 99, 162 e 286). Esses res_i elevados estão relacionados, principalmente, a localização das observações, a maioria com valores de ELEV ao redor de 300 m, zona de transição entre os domínios fisiográficos estudados (Figura 8).

Quanto às observações mais influenciais, três são provenientes de solos derivados de rochas ígneas extrusivas (observações nº 140, 162 e 302) e três de solos derivados de rochas e materiais sedimentares (observações nº 1, 166 e 173). Note que a observação nº 162 possui res_i elevado, além de também ter sido influencial na FPESe descrita acima, assim como as observações de nº140 e 302. A influência dessas observações pode ser devida ao valores absolutos de CONV e IPE elevados.

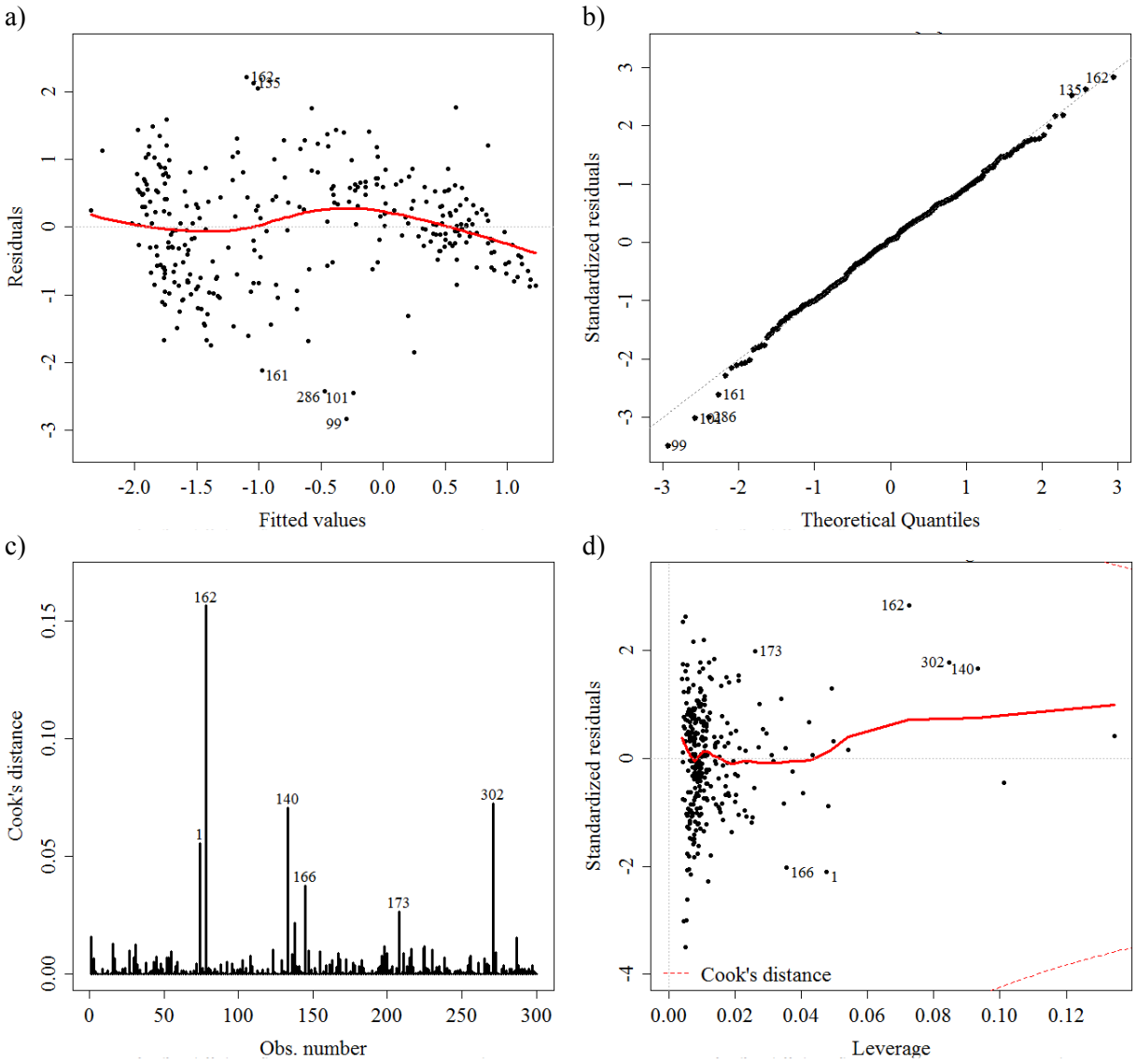


Figura 21 – Análise gráfica da qualidade do ajuste da função de predição espacial da log-razão aditiva ln(silte/areia) em toda a área de estudo (n = 300): a) valores preditos (*fitted values*) e resíduos (*residuals*), b) quantis teóricos (*theoretical quantiles*) e resíduos padronizados (*standardized residuals*), c) distância de Cook (*Cook's distance*) associada a cada observação (*obs. number*), d) alavancagem (*leverage*) e resíduos padronizados (*standardized residuals*).

Tabela 16 – Identificação das observações atípicas e influenciais na função de predição espacial da log-razão aditiva ln(silte/areia) em toda a área de estudo.

ID	Material de origem	Classe de solo	Areia	Silte	Argila	CONV	ELEV	IPE
1	Sedimentar	RF	93	3	4	-38,2813	149,5288	9,063178
99	Sedimentar	RQ	91	4	5	-3,9302	342,751	6,123335
101	Sedimentar	RL	88	6	6	-0,53634	350,3945	5,692258
135	Ígnea	RL	20	59	21	4,329094	268,4856	5,444137
140	Ígnea	RL	26	42	32	54,02481	341,8192	1,216597
161	Sedimentar	RR	88	4	8	-1,45345	277,6776	6,579843
162	Ígnea	RL	17	52	31	-48,3327	227,5129	11,18767
166	Sedimentar	RL	89	6	5	-23,0368	273,7695	10,94474
173	Sedimentar	PBAC	44	38	18	0,030519	148,1155	2,2123
286	Sedimentar	RQ	91	5	4	1,077544	332,2042	6,06097
302	Ígnea	RL	21	53	26	-52,382	283,0784	10,51904

Nota: ID – identificação da observação, RL – Neossolo Litólico, RQ – Neossolo Quartzarênico, RF – Neossolo Flúvico, RR – Neossolo Regolítico, PBAC – Argissolo Bruno-Acinzentado, CONV – índice de convergência (%), ELEV – elevação (m), IPE – logaritmo natural do índice de potência de escoamento (adimensional).

As estatísticas da validação cruzada (339 observações, 10 partições, 100 repetições) mostram que as FPESe construídas para estimar as log-razões aditivas ln(argila/areia) e ln(silte/areia) são capazes de explicar entre 50 e 75% da variância da distribuição do tamanho de partículas do solo na área de estudo (Tabela 17). As melhores predições são as da fração areia, para a qual as FPESe são capazes de explicar mais de 70% da variância (R^2_{aj} = 0,74). As predições mais pobres são aquelas da fração argila (R^2_{aj} = 0,51), ficando as predições da fração silte em posição intermediária (R^2_{aj} = 0,68). Apesar de os erros de predição (*RMSE*) serem maiores para a fração areia, e menores para a fração argila, proporcionalmente ($RMSE_{norm}$), os maiores erros de predição são os da fração silte ($RMSE_{norm}$ = 0,18). Quando avaliado o erro médio (*EM*) das predições, pode-se perceber que a fração areia é superestimada em 1,44%, enquanto argila e silte são subestimados em 0,43% e 1,01%, respectivamente.

A Figura 22 mostra os valores preditos plotados contra os valores medidos das três frações de tamanho de partícula do solo. A dispersão dos pontos ao redor da linha 1:1 é acentuada, sendo mais pronunciada em valores intermediários. Além disso, nos três gráficos (a, b e c) é possível notar a formação de dois grupos de pontos, comportamento que está

intimamente relacionado ao material de origem do solo (Figura **9**). Observações provenientes de solos derivados de rochas ígneas extrusivas possuem distribuição do tamanho de partículas mais fina do que solos derivados de rochas e materiais sedimentares.

Tabela 17 – Estatísticas da validação cruzada (339 observações; 10 partições; 100 repetições) das funções de predição espacial da distribuição do tamanho de partículas em toda a área de estudo.

Fração de tamanho de partícula	Estatísticas[1]	Percentil 2,5	Mediana	Percentil 97,5
Areia	*RMSE* (%)	14,3676369	14,4126949	14,4902126
	$RMSE_{norm}$	0,1651453	0,1656632	0,1665542
	EM (%)	1,3985053	1,4379407	1,4892336
	R^2_{aj}	0,7387924	0,7406735	0,7437131
Silte	*RMSE* (%)	11,5181723	11,5905996	11,6812441
	$RMSE_{norm}$	0,1772027	0,1783169	0,1797114
	EM (%)	-1,0571834	-1,0117448	-0,9666558
	R^2_{aj}	0,6773583	0,6828690	0,6892626
Argila	*RMSE* (%)	6,7909921	6,8371745	6,9121440
	$RMSE_{norm}$	0,1331567	0,1340622	0,1355322
	EM (%)	-0,4704550	-0,4267001	-0,3825939
	R^2_{aj}	0,5080010	0,5148402	0,5218994

[1] *RMSE* – raiz quadrada do erro quadrático médio, $RMSE_{norm}$ – raiz quadrado do erro quadrático médio normalizada, *EM* – erro médio, R^2_{aj} - coeficiente de determinação ajustado.

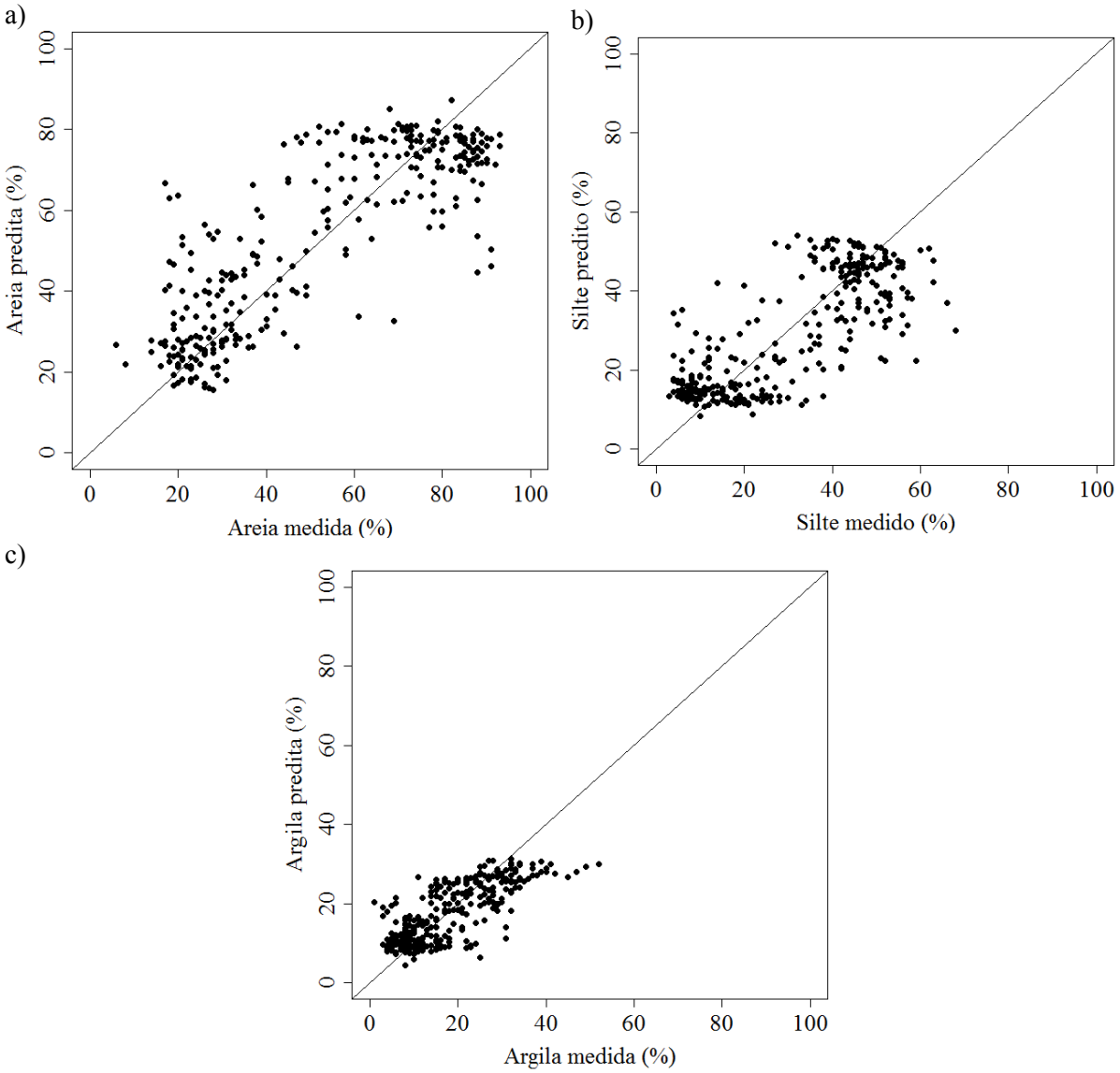


Figura 22 – Distribuição do tamanho de partícula medida e predita pelas funções de predição espacial da distribuição do tamanho de partículas do solo em toda a área de estudo (n = 300).

6.2.2 FPESe para os domínios fisiográficos

6.2.2.1 Domínio fisiográfico inferior

A FPESe que construí para estimar a log-razão aditiva ln(argila/areia) no domínio fisiográfico inferior (ELEV < 300 m) possui em sua estrutura as mesmas variáveis preditoras

(ELEV, IPE e CONV) que utilizei para construir as FPESe apresentadas acima (Tabela 18). Entretanto, a FPESe construída

$$\ln\left(\frac{argila}{areia}\right) = -2{,}660236 - 0{,}021394CONV + 0{,}005210ELEV - 0{,}073006IPE\,, \qquad (38)$$

onde CONV é o índice de convergência (%), ELEV é a elevação (m) e IPE é o logaritmo natural do índice de potência de escoamento, explica apenas 12% da variância (estimada pela soma de quadrados total) (R^2 = 0,1193) (Tabela 18 e Tabela 19). A maior parte da variância está contida nos res_i ($SQ_{resíduos}$ = 65,979 → 88% da variância).

Tabela 18 – Coeficientes da função de predição espacial da log-razão aditiva ln(argila/areia) no domínio fisiográfico inferior (n = 150).

ln(argila/areia)	Estimativa	Erro padrão	Valor t	Pr(>|t|)
Intercepto	-2,660236	0,275018	-9,673	< 2e-16
CONV[1]	-0,021394	0,006383	-3,352	0,001022
ELEV	0,005210	0,001518	3,432	0,000781
IPE	-0,073006	0,036704	-1,989	0,048567
Erro quadrático médio: 0,6722 com 146 graus de liberdade				
R² múltiplo: 0,1193; R² ajustado: 0,1012				
Critério de Informação de Akaike: -115,1958				

[1] Variáveis preditoras: ELEV – elevação, CONV – índice de convergência, IPE – logaritmo natural do índice de poder de escoamento.

Os coeficientes da FPESe construída para estimar a log-razão aditiva ln(argila/areia) também indicam que a proporção da fração argila aumenta com o aumento de ELEV (coeficiente com sinal positivo), enquanto a proporção da fração areia aumenta com o aumento de IPE e CONV (coeficiente com sinal negativo) (Tabela 18).

As variáveis que mais contribuem para a explicação da variância são ELEV (EV = 5%) e CONV (EV = 5%), enquanto IPE explica uma pequena parte da variância (EV = 1%), conforme calculado através da Equação (21). Note que, diferente do que ocorreu para as

FPESe construídas para toda a área de estudo, agora CONV apresentou maior importância do que IPE (Tabela 19).

Tabela 19 – Análise de variância da função de predição espacial da log-razão aditiva ln(argila/areia) no domínio fisiográfico inferior (*n* = 150).

Fonte de variação	Graus de liberdade	Soma de quadrados	Quadrado médio	F	Pr(>F)
Regressão[1]	3	8,939	2,9796	6,594	0,0003286
CONV	1	3,583	3,5831	7,9288	0,005538
ELEV	1	3,568	3,5682	7,8958	0,005635
IPE	1	1,788	1,7878	3,9562	0,048567
Resíduos	146	65,979	0,4519	-	-
Total	149	74,918	0.5028	-	-

[1] Variáveis preditoras: ELEV – elevação, CONV – índice de convergência, IPE – logaritmo natural do índice de poder de escoamento.

Seguindo o padrão observado até então, as variáveis preditoras incluídas na FPESe são pouco correlacionadas: a variância dos parâmetros do modelo de regressão ajustado é pequena e o *FIV* é inferior a 2,0 (Tabela 20). Isso significa, mais uma vez, que não existe multicolinearidade no conjunto de variáveis preditoras utilizado (ELEV, CONV e IPE).

Tabela 20 – Matriz variância-covariância dos parâmetros ajustados da função de predição espacial da log-razão aditiva ln(argila/areia) no domínio fisiográfico inferior (*n* = 150) e fator de inflação da variância (FIV) associado a cada variável preditora.

	Intercepto	CONV	ELEV	IPE	FIV
Intercepto	0,0756348361	-1,331142e-04	-3,162353e-04	-1,483460e-03	-
CONV[1]	-0,0001331142	4,074225e-05	-2,429459e-06	1,279192e-04	1,426898
ELEV	-0,0003162353	-2,429459e-06	2,305461e-06	-2,844325e-05	1,354262
IPE	-0,0014834598	1,279192e-04	-2,844325e-05	1,347210e-03	1,808235

[1] Variáveis preditoras: ELEV – elevação, CONV – índice de convergência, IPE – logaritmo natural do índice de poder de escoamento.

Os res_i das predições não são relacionados aos valores preditos, sendo a maioria dos r_i < |2,0|. Apesar de haverem res_i^* > 1,96 e h_{ii} > $3p/n$ = 3×4/150 = 0,08, não removi nenhuma observação do conjunto de dados para o ajuste dos modelos de regressão. Mais uma vez me pautei pelos valores de D_i, que foram todos < 1,0, ou seja, abaixo do limite proposto por Weisberg (2005).

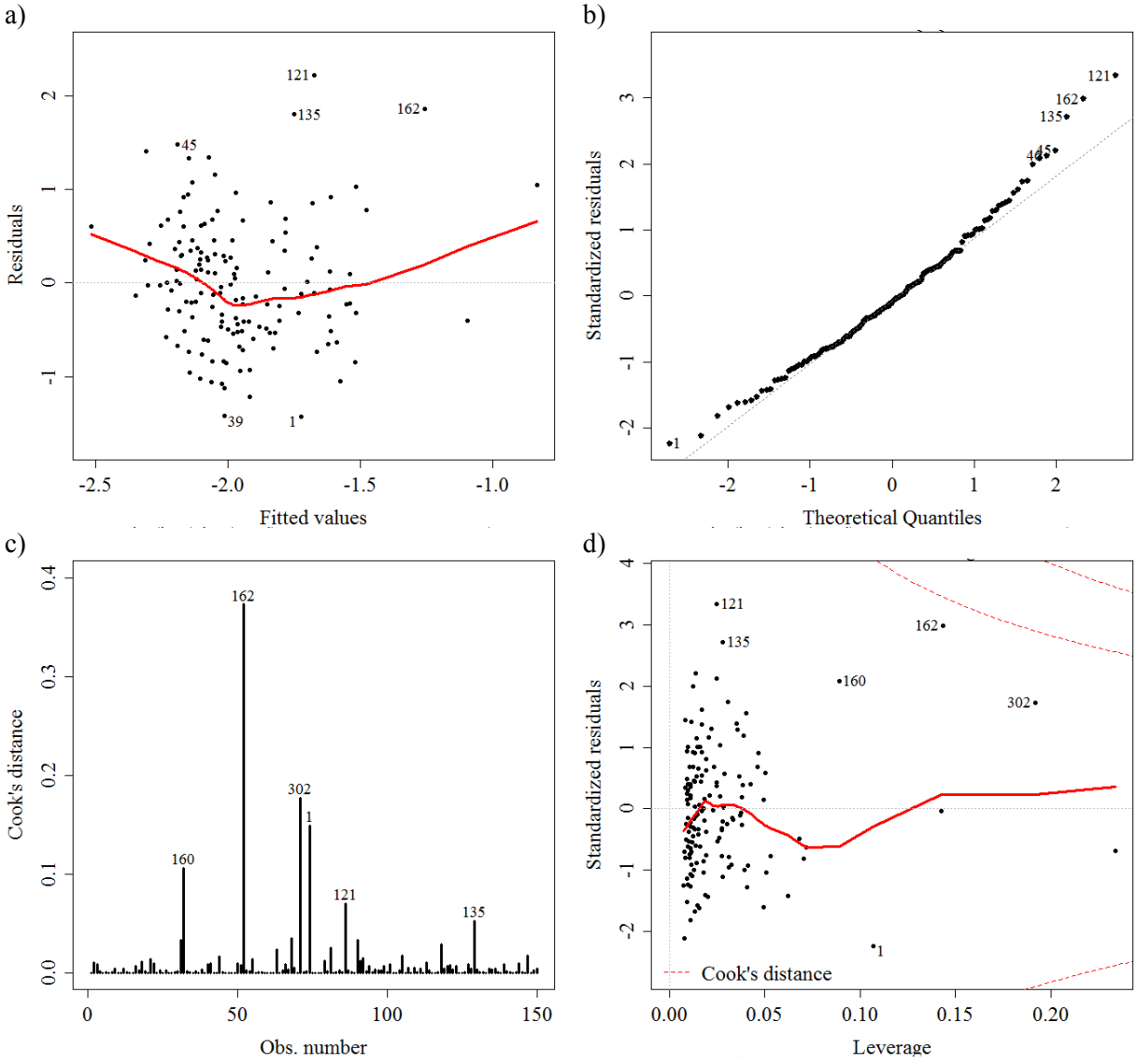


Figura 23 – Análise gráfica da qualidade do ajuste da função de predição espacial da log-razão aditiva ln(argila/areia) no domínio fisiográfico inferior (n = 150): a) valores preditos (*fitted values*) e resíduos (*residuals*), b) quantis teóricos (*theoretical quantiles*) e resíduos padronizados (*standardized residuals*), c) distância de Cook (*Cook's distance*) associada a cada observação (*obs. number*), d) alavancagem (*leverage*) e resíduos padronizados (*standardized residuals*).

Dentre as seis observações que apresentaram os maiores res_i, três são provenientes de solos derivados de rochas e materiais sedimentares (observações nº 1, 39 e 45) e três são provenientes de solos derivados de rochas ígneas extrusivas (observações nº 121, 135 e 162) (Tabela 21). Note que as três observações provenientes de solos derivados de rochas ígneas extrusivas já apresentaram res_i significativos nas FPESe avaliadas anteriormente. Dado que o domínio fisiográfico inferior é constituído, predominantemente, por solos derivados de rochas e materiais sedimentares, as observações provenientes de solos derivados de rochas ígneas extrusivas estão localizadas logo abaixo da cota de 300 m, definida como limítrofe dos domínios fisiográficos estudados. No que diz respeito às observações provenientes de solos derivados de rochas e materiais sedimentares, duas delas (observações nº 1 e 39) possuem distribuição do tamanho de partículas muito similar, mas valores das variáveis preditoras bastante diferentes. Já na terceira (observação nº 45) a proporção de cada uma das frações de tamanho de partícula é similar, característica incomum à maioria dos solos derivados de rochas e materiais sedimentares na área de estudo.

Tabela 21 – Identificação das observações atípicas e influenciais na função de predição espacial da log-razão aditiva ln(argila/areia) no domínio fisiográfico inferior.

ID	Material de origem	Classe de solo	Areia	Silte	Argila	CONV	ELEV	IPE
1	Sedimentar	RF	93	3	4	-38,2813	149,5288	9,063178
39	Sedimentar	PBAC	93	4	3	-2,90269	185,0437	5,214863
45	Sedimentar	PBAC	47	30	23	6,592047	178,202	4,363
121	Ígnea	RL	18	51	31	-1,16179	268,1029	5,963155
135	Ígnea	RL	20	59	21	4,329094	268,4856	5,444137
160	Ígnea	RL	52	23	25	-15,8443	216,2039	12,02097
162	Ígnea	RL	17	52	31	-48,3327	227,5129	11,18767
302	Ígnea	RL	21	53	26	-52,382	283,0784	10,51904

Nota: ID – identificação da observação, RL – Neossolo Litólico, RF – Neossolo Flúvico, PBAC – Argissolo Bruno-Acinzentado, CONV – índice de convergência (%), ELEV – elevação (m), IPE – logaritmo natural do índice de potência de escoamento (adimensional).

Com relação às seis observações mais influentes, cinco delas (observações nº 121, 135, 160, 162 e 302) são provenientes de solos derivados de rochas ígneas extrusivas (as observações nº 160, 162 e 302 também estão entre as mais influentes nas FPESe descritas acima). Em um domínio fisiográfico onde predominam solos derivados de rochas e materiais

sedimentares tal resultado pode ser considerado dentro da normalidade. Já a observação proveniente de solo derivado de material sedimentar é aquela de nº 1, a qual já apresentou res_i significativo nas FPESe descritas acima. Trata-se de uma amostra de solo Neossolo Flúvico formado a partir de depósitos fluviais recentes.

Já a FPESe que construí para estimar a log-razão aditiva ln(silte/areia) no domínio fisiográfico inferior (ELEV < 300 m) é um pouco diferente das demais FPESe que mostrei até agora. Ela possui em sua estrutura as mesmas variáveis preditoras (ELEV, IPE e CONV) (Tabela 22). A proporção da variância explicada pela FPESe construída

$$\ln\left(\frac{silte}{areia}\right) = -2{,}143823 - 0{,}032446CONV + 0{,}006064ELEV - 0{,}136437IPE\,, \qquad (39)$$

onde CONV é o índice de convergência (%), ELEV é a elevação (m) e IPE é o logaritmo natural do índice de potência de escoamento, também é pequena. Ela explica apenas 13% da variância (estimada pela soma de quadrados total) (R_2 = 0,1279) (Tabela 22 e Tabela 23). Mas a maior diferença está na proporção da variância explicada pelas variáveis preditoras. As variáveis que mais contribuem para a explicação da variância são CONV (EV = 5%) e IPE (EV = 5%), enquanto ELEV, que sempre foi a mais importante, explica uma pequena parte da variância (EV = 2%).

Tabela 22 – Coeficientes da função de predição espacial da log-razão aditiva ln(silte/areia) no domínio fisiográfico inferior (n = 150).

ln(silte/areia)	Estimativa	Erro padrão	Valor t	Pr(>|t|)
Intercepto	-2,143823	0,345320	-6,208	5,27e-09
CONV[1]	-0,032446	0,008015	-4,048	8,34e-05
ELEV	0,006064	0,001907	3,181	0,00179
IPE	-0,136437	0,046087	-2,960	0,00359
Erro quadrático médio: 0,8441 com 146 graus de liberdade				
R² múltiplo: 0,1279; R² ajustado: 0,11				
Critério de Informação de Akaike: -46,90547				

[1] Variáveis preditoras: ELEV – elevação, CONV – índice de convergência, IPE – logaritmo natural do índice de poder de escoamento.

Os coeficientes da FPESe construída para estimar a log-razão aditiva ln(silte/areia) também indicam que a proporção da fração silte aumenta com o aumento de ELEV (coeficiente com sinal positivo), enquanto a proporção da fração areia aumenta com o aumento de IPE e CONV (coeficiente com sinal negativo) (Tabela 22).

Tabela 23 – Análise de variância da função de predição espacial da log-razão aditiva ln(silte/areia) no domínio fisiográfico inferior (*n* = 150).

Fonte de variação	Graus de liberdade	Soma de quadrados	Quadrado médio	F	Pr(>F)
Regressão[1]	3	15,26	5,0866	7,14	0,0001655
CONV	1	6,329	6,3293	8,8835	0,003372
ELEV	1	2,687	2,687	3,7710	0,054074
IPE	1	6,244	6,244	8,7641	0,003586
Resíduos	146	104,022	0,7125	-	-
Total	149	119,282	0,8005	-	-

[1] Variáveis preditoras: ELEV – elevação, CONV – índice de convergência, IPE – logaritmo natural do índice de poder de escoamento.

Assim como nas demais FPESe construídas, as variáveis preditoras são pouco correlacionada (Tabela 24), mostrando que não há multicolinearidade na FPESe construída.

Tabela 24 - Matriz variância-covariância dos parâmetros ajustados da função de predição espacial da log-razão aditiva ln(silte/areia) no domínio fisiográfico inferior (*n* = 150)e fator de inflação da variância (FIV) associado a cada variável preditora.

	Intercepto	CONV	ELEV	IPE	FIV
Intercepto	0,1192456520	-2,098675e-04	-4,985757e-04	-2,338818e-03	-
CONV[1]	-0,0002098675	6,423411e-05	-3,830278e-06	2,016770e-04	1,426898
ELEV	-0,0004985757	-3,830278e-06	3,634783e-06	-4,484353e-05	1,354262
IPE	-0,0023388181	2,016770e-04	-4,484353e-05	2,124008e-03	1,808235

[1] Variáveis preditoras: ELEV – elevação, CONV – índice de convergência, IPE – logaritmo natural do índice de poder de escoamento.

Novamente, utilizado os critérios de avaliação das observações atípicas e influentes descritos acima, nenhuma observação foi eliminada do conjunto de dados para o ajuste dos modelos de regressão. Apesar de haverem observações com $res_i > |2,0|$, $res_i^* > 1,96$ e $h_{ii} > 3p/n = 3\times4/150 = 0,08$, não houveram valores de D_i superiores a 1,0. Além disso, os res_i não são relacionados aos valores preditos (Figura 24).

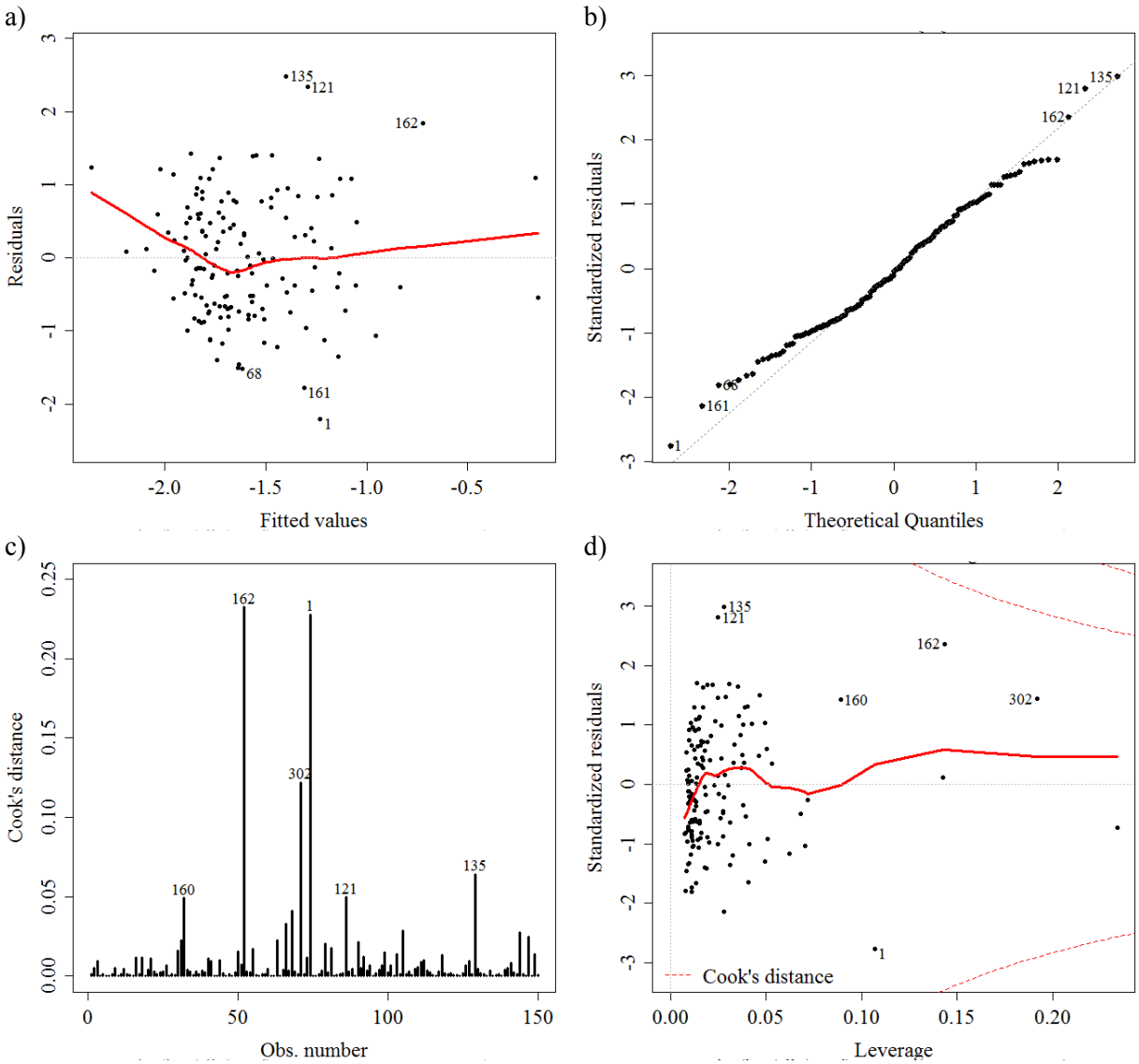


Figura 24 – Análise gráfica da qualidade do ajuste da função de predição espacial da log-razão aditiva ln(silte/areia) no domínio fisiográfico inferior (n = 150): a) valores preditos (*fitted values*) e resíduos (*residuals*), b) quantis teóricos (*theoretical quantiles*) e resíduos padronizados (*standardized residuals*), c) distância de Cook (*Cook's distance*) associada a cada observação (*obs. number*), d) alavancagem (*leverage*) e resíduos padronizados (*standardized residuals*).

Dentre as seis observações que apresentaram os maiores res_i, três são provenientes de solos derivados de rochas e materiais sedimentares (observações nº 1, 68 e 161) e três são provenientes de solos derivados de rochas ígneas extrusivas (observações nº 162, 135 e 121) (Tabela 25). De todas as seis, apenas uma observação (nº 68) não apresentou res_i elevado nas FPESe descritas acima. Dado que o domínio fisiográfico inferior é constituído, predominantemente, por solos derivados de rochas e materiais sedimentares, é normal que as observações provenientes de solos derivados de rochas ígneas extrusivas apresentem res_i elevados. No que diz respeito às observações provenientes de solos derivados de rochas e materiais sedimentares, duas delas (observações nº 68 e 161) localizam-se próximo da cota de 300 m. A terceira é aquela de nº 1, já discutida acima.

Tabela 25 – Identificação das observações atípicas e influenciais na função de predição espacial da log-razão aditiva ln(silte/areia) no domínio fisiográfico inferior.

ID	Material de origem	Classe de solo	Areia	Silte	Argila	CONV	ELEV	IPE
1	Sedimentar	RF	93	3	4	-38,2813	149,5288	9,063178
68	Sedimentar	PBAC	92	4	4	0,589827	234,4887	6,408733
121	Ígnea	RL	18	51	31	-1,16179	268,1029	5,963155
135	Ígnea	RL	20	59	21	4,329094	268,4856	5,444137
160	Ígnea	RL	52	23	25	-15,8443	216,2039	12,02097
161	Sedimentar	RR	88	4	8	-1,45345	277,6776	6,579843
162	Ígnea	RL	17	52	31	-48,3327	227,5129	11,18767
302	Ígnea	RL	21	53	26	-52,382	283,0784	10,51904

Nota: ID – identificação da observação, RL – Neossolo Litólico, RR – Neossolo Regolítico, PBAC – Argissolo Bruno-Acinzentado, CONV – índice de convergência (%), ELEV – elevação (m), IPE – logaritmo natural do índice de potência de escoamento (adimensional).

Com relação às seis observações mais influentes, a maioria delas (observações nº 121, 135, 160, 162 e 302) são provenientes de solos derivados de rochas ígneas extrusivas, mais a observação de nº 1, proveniente de solo desenvolvido a partir de depósitos fluviais recentes. Todas essas observações se mostraram as mais influentes na FPESe construída para estimar a log-razão aditiva ln(argila/areia) no domínio fisiográfico inferior (Tabela 25).

As estatísticas da validação cruzada (165 observações, 10 partições, 100 repetições) mostram que as FPESe construídas para estimar as log-razões aditivas ln(argila/areia) e ln(silte/areia) são capazes de explicar entre 8% e 13% da variância da distribuição do tamanho

de partículas do solo (Tabela 26). As melhores predições são as da fração silte, para a qual a FPESe é capaz de explicar 13% da variância (R^2_{aj} = 0,13). As predições mais pobres são novamente aquelas da fração argila (R^2_{aj} = 0,08), ficando as predições da fração areia em posição intermediária (R^2_{aj} = 0,12). Note que os erros de predição (*RMSE*) são maiores para a fração areia, mas quando avaliados proporcionalmente ($RMSE_{norm}$), todos alcançam o mesmo valor ($RMSE_{norm}$ = 0,19). Quanto aos erros médios (*EM*) de predição, mais uma vez a fração areia é sobrestimada (*EM* = 2,24%), enquanto as frações argila (*EM* = -0,41%) e silte (*EM* = -1,84%) são subestimadas.

Tabela 26 – Estatísticas da validação cruzada (165 observações; 10 partições; 100 repetições) das funções de predição espacial da distribuição do tamanho de partículas no domínio fisiográfico inferior.

Fração de tamanho de partícula	Estatísticas[1]	Percentil 2,5	Mediana	Percentil 97,5
Areia	*RMSE* (%)	14,8698323	15,0710391	15,2761450
	$RMSE_{norm}$	0,1956557	0,1983031	0,2010019
	EM (%)	2,1337222	2,2367254	2,3457734
	R^2_{aj}	0,1088800	0,1200910	0,1368903
Silte	*RMSE* (%)	10,4755348	10,6722087	10,9794385
	$RMSE_{norm}$	0,1870631	0,1905752	0,1960614
	EM (%)	-1,9622973	-1,8398853	-1,7081304
	R^2_{aj}	0,1098410	0,1287955	0,1661061
Argila	*RMSE* (%)	5,24556866	5,32742190	5,4891449
	$RMSE_{norm}$	0,18734174	0,19026507	0,1960409
	EM (%)	-0,46742132	-0,40797474	-0,3461830
	R^2_{aj}	0,07081797	0,08327774	0,1109848

[1] *RMSE* – raiz quadrada do erro quadrático médio, $RMSE_{norm}$ – raiz quadrado do erro quadrático médio normalizada, *EM* – erro médio, R^2_{aj} - coeficiente de determinação ajustado.

A Figura 25 mostra os valores preditos plotados contra os valores medidos das três frações de tamanho de partícula (n = 165). A dispersão dos pontos é acentuada e não segue a tendência da linha 1:1. Confirmando as estatísticas da validação cruzada (Tabela 26), a proporção da fração areia é sobrestimada (pontos a esquerda da linha 1:1), enquanto a proporção das frações silte e argila é subestimada (pontos a direita da linha 1:1).

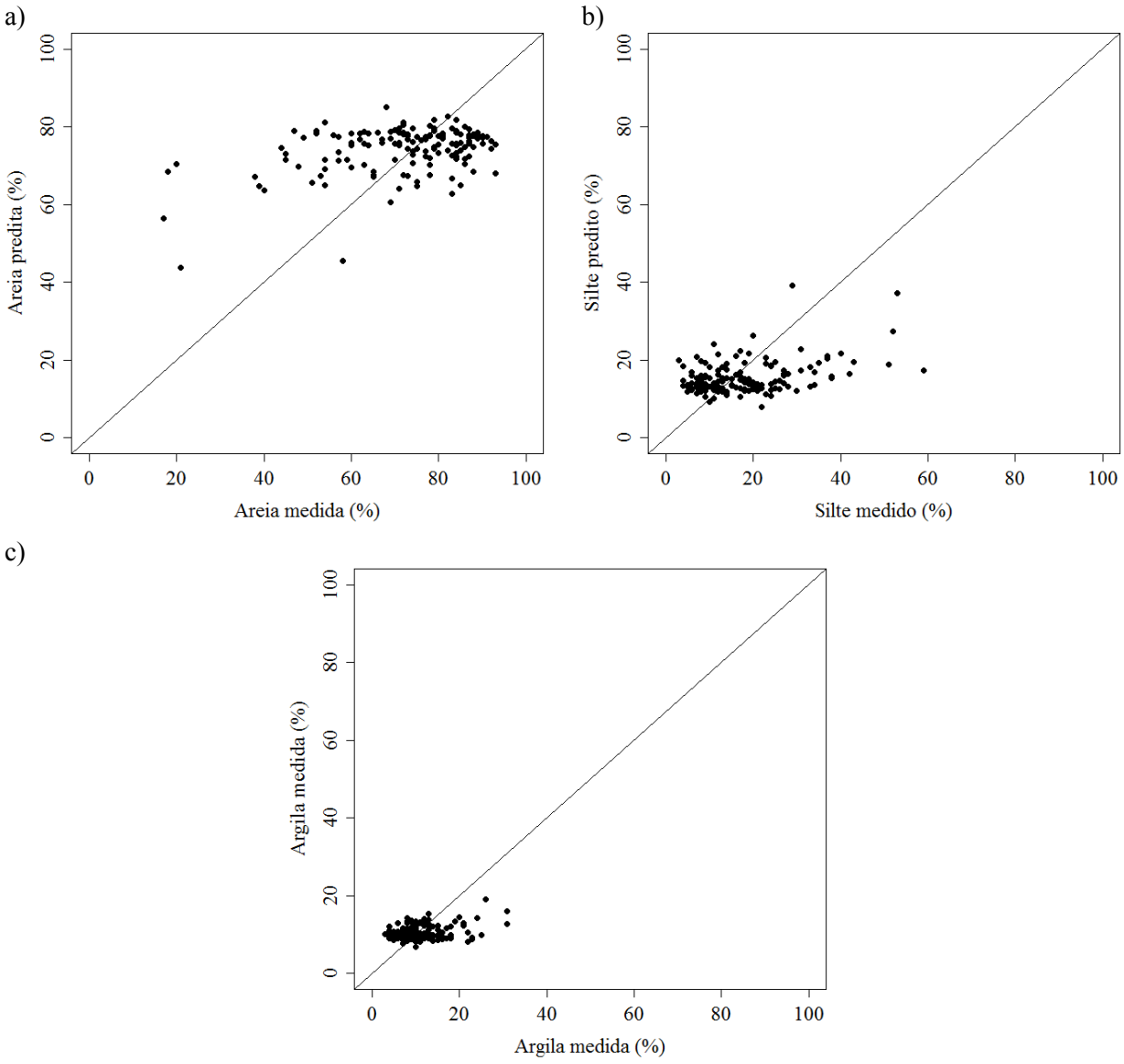


Figura 25 – Distribuição do tamanho de partícula (a – areia, b – silte, c – argila) medida e predita pelas funções de predição espacial da distribuição do tamanho de partículas do solo no domínio fisiográfico inferior (n = 150).

6.2.2.2 Domínio fisiográfico superior

A FPESe que construí para estimar a log-razão aditiva ln(argila/areia) no domínio fisiográfico superior (ELEV ≥ 300 m) possui estrutura significativamente diferente das FPESe construídas para estimar as log-razões aditivas em toda a área de estudo e no domínio fisiográfico inferior. Através do procedimento *stepwise*, tendo como critério de seleção das

variáveis a minimização do *AIC*, apenas as variáveis preditoras ELEV e IPE foram incluídas no modelo de regressão ajustado (Tabela 27 e Tabela 28). Essa FPESe construída

$$\ln\left(\frac{argila}{areia}\right) = -3{,}792363 + 0{,}009920\,ELEV - 0{,}106727\,IPE\,, \qquad (40)$$

onde ELEV é a elevação (m) e IPE é o logaritmo natural do índice de potência de escoamento (adimensional), explica 42% da variância (estimada pela soma de quadrados total) (R^2 = 0,4201) (Tabela 27 e Tabela 28). Seguindo o mesmo padrão das demais FPESe construídas, os coeficientes da FPESe construída para estimar a log-razão aditiva ln(argila/areia) indicam que a proporção da fração argila aumenta com o aumento de ELEV (coeficiente com sinal positivo), enquanto a proporção da fração areia aumenta com o aumento de IPE (coeficiente com sinal negativo) (Tabela 27).

Tabela 27 – Coeficientes da função de predição espacial da log-razão aditiva ln(argila/areia) no domínio fisiográfico superior (*n* = 150).

ln(argila/areia)	Estimativa	Erro padrão	Valor t	Pr(>|t|)
Intercepto	-3,792363	0,725064	-5,230	5,72e-07
ELEV[1]	0,009920	0,001477	6,716	3,82e-10
IPE	-0,106727	0,042553	-2,508	0,0132
Erro quadrático médio: 0,7182 com 147 graus de liberdade				
R^2 múltiplo: 0,4201; R^2 ajustado: 0,4122				
Critério de Informação de Akaike: -96,33882				

[1] Variáveis preditoras: ELEV – elevação, IPE – logaritmo natural do índice de poder de escoamento.

Assim como para as FPESe construídas para estimar as log-razões aditivas em toda a área de estudo, a variável que mais contribui para a explicação da variância é ELEV (*EV* = 39%), enquanto IPE explica apenas uma pequena parte da variância (*EV* = 3%) (Tabela 28), conforme calculado através da Equação (21). Essas variáveis são pouco correlacionadas, mostrando que não há ocorrência de multicolinearidade na FPESe construída (Tabela 29).

Todos os FIV ficaram abaixo de 2,0 e a variância dos coeficientes do modelo de regressão é reduzida.

Tabela 28 – Análise de variância da função de predição espacial da log-razão aditiva ln(argila/areia) no domínio fisiográfico superior (n = 150).

Fonte de variação	Graus de liberdade	Soma de quadrados	Quadrado médio	F	Pr(>F)
Regressão[1]	2	54,918	27,459	53,24	< 2,2e-16
ELEV	1	51,673	51,673	100,1814	< 2e-16
IPE	1	3,245	3,245	6,2907	0,01322
Resíduos	147	75,821	0,516	-	-
Total	149	130,739	0.8774	-	-

[1] Variáveis preditoras: ELEV – elevação, IPE – logaritmo natural do índice de poder de escoamento.

Tabela 29 – Matriz variância-covariância dos parâmetros ajustados da função de predição espacial da log-razão aditiva ln(argila/areia) no domínio fisiográfico superior (n = 150) e fator de inflação da variância (FIV) associado a cada variável preditora.

	Intercepto	ELEV	IPE	FIV
Intercepto	0,525717959	-1,040733e-03	-2,291789e-02	-
ELEV[1]	-1,040733e-03	2,181788e-06	3,634559e-05	1,502356
IPE	-2,291789e-02	3,634559e-05	1,810722e-03	1,502356

[1] Variáveis preditoras: ELEV – elevação, IPE – logaritmo natural do índice de poder de escoamento.

Mais uma vez há observações com $res_i > |2,0|$, $res_i^* > 1,96$ e $h_{ii} > 3p/n = 3\times4/150 = 0,08$ (Figura 26). Mesmo assim, conforme já procedi anteriormente, não removi nenhuma observação do conjunto de dados para o ajuste dos modelos de regressão. Todos os valores de D_i foram < 1,0, ou seja, abaixo do limite proposto por Weisberg (2005). Além disso, os res_i não são relacionados aos valores preditos.

Dentre as seis observações que apresentaram os maiores res_i, cinco são provenientes de solos derivados de rochas e materiais sedimentares (observações nº 99, 101, 116, 117 e 328) e uma é proveniente de solo derivado de rocha ígnea extrusiva (observação nº 226) (Tabela 30). Quatro delas já apresentaram res_i elevados nas FPESe descritas acima (observações nº 99, 101, 116 e 328). Dado que o domínio fisiográfico superior é constituído,

predominantemente, por solos derivados de rochas ígneas extrusivas, as observações provenientes de solos derivados de rochas e materiais sedimentares estão localizadas logo acima da cota de 300 m, definida como limítrofe dos domínios fisiográficos estudados. No que diz respeito à observação proveniente de solo derivado de rocha ígnea extrusiva, aquela de nº 226 possui IPE relativamente inferior se comparada às demais observações.

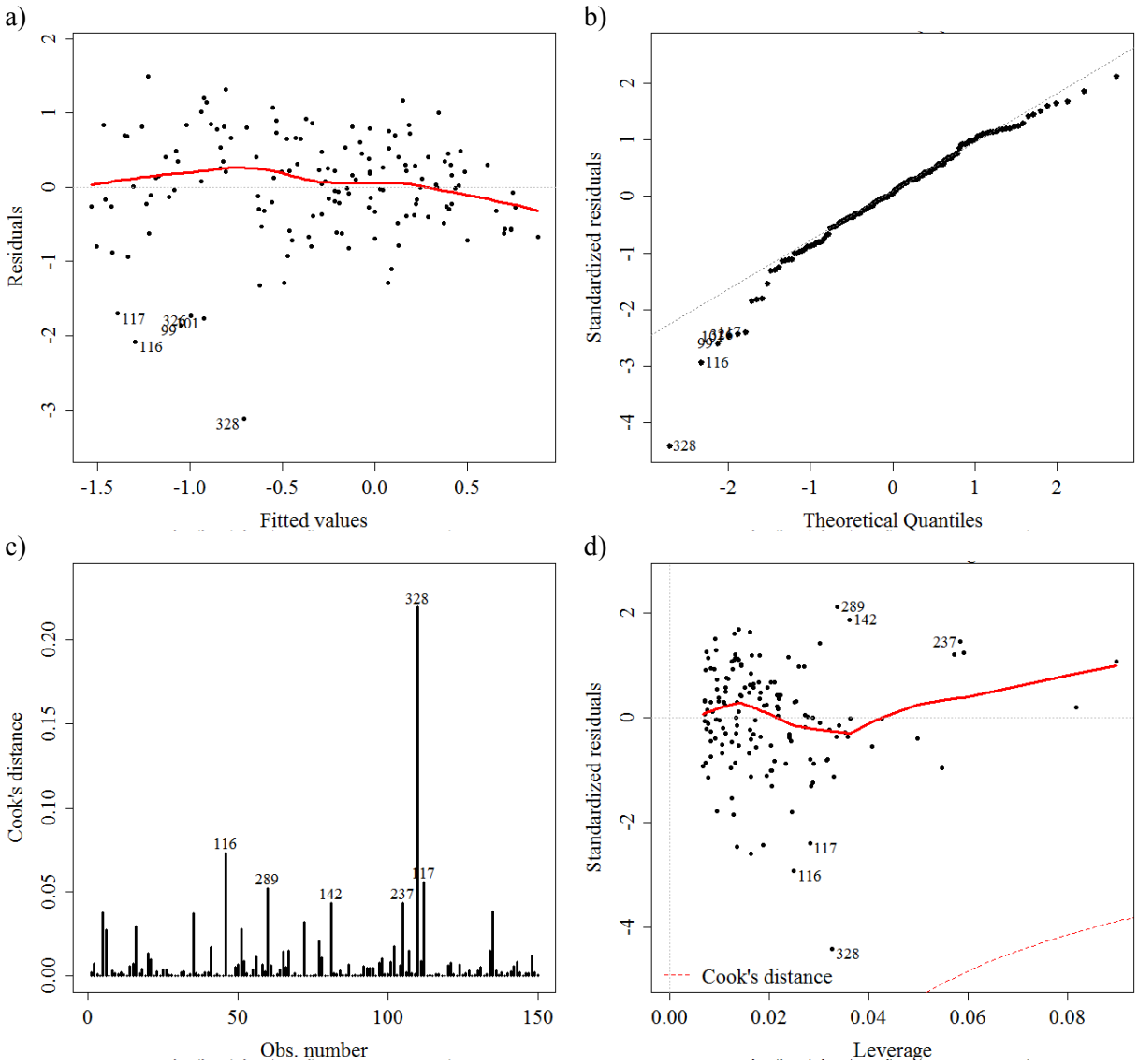


Figura 26 – Análise gráfica da qualidade do ajuste da função de predição espacial da log-razão aditiva ln(argila/areia) no domínio fisiográfico superior: a) valores preditos (*fitted values*) e resíduos (*residuals*), b) quantis teóricos (*theoretical quantiles*) e resíduos padronizados (*standardized residuals*), c) distância de Cook (*Cook's distance*) associada a cada observação (*obs. number*), d) alavancagem (*leverage*) e resíduos padronizados (*standardized residuals*).

Com relação às seis observações mais influentes, três delas (observações nº 116, 117 e 328) são provenientes de solos derivados de rochas e materiais sedimentares. As outras três (observações nº 142, 237 e 289) são provenientes de solos derivados de rochas ígneas extrusivas. Das seis, apenas a observação nº 328 se mostrou influente em nas FPESe descritas acima.

Tabela 30 – Identificação das observações atípicas e influenciais na função de predição espacial da log-razão aditiva ln(argila/areia) no domínio fisiográfico superior.

ID	Material de origem	Classe de solo	Areia	Silte	Argila	CONV	ELEV	IPE
99	Sedimentar	RQ	91	4	5	-3,9302	342,751	6,123335
101	Sedimentar	RL	88	6	6	-0,53634	350,3945	5,692258
116	Sedimentar	PV	88	9	3	0,518997	319,407	6,30546
142	Ígnea	RL	18	52	30	14,44046	339,0061	3,52416
226	Ígnea	RL	18	53	29	20,79176	450,4106	2,106457
237	Ígnea	RL	27	44	29	24,17803	320,9496	3,088782
328	Ígnea	CX	46	53	1	-2,1352	391,5299	7,508506

Nota: ID – identificação da observação, RL – Neossolo Litólico, RQ – Neossolo Quartzarênico, PV – Argissolo Vermelho, CX – Cambissolo Háplico, CONV – índice de convergência (%), ELEV – elevação (m), IPE – logaritmo natural do índice de potência de escoamento (adimensional).

A FPESe que construí para estimar a log-razão aditiva ln(silte/areia) no domínio fisiográfico superior (ELEV ≥ 300 m) é a mais simples de todas as FPESe construídas. Isso porque ela possui em sua estrutura apenas a variável preditora ELEV, selecionada através do procedimento *stepwise* tendo como critério a minimização do *AIC* (Tabela 31 e Tabela 32). A FPESe construída

$$\ln\left(\frac{silte}{areia}\right) = -3{,}067589 + 0{,}008448 ELEV \ , \qquad (41)$$

onde ELEV é a elevação (m), explica 26% da variância (estimada pela soma de quadrados total) (R_2 = 0,2668) (Tabela 31 e Tabela 32), variância essa explicada em sua totalidade pela variável preditora ELEV. Os coeficientes da FPESe também indicam que a proporção da

fração silte aumenta com o aumento de ELEV (coeficiente com sinal positivo), enquanto a proporção da fração areia diminui com o aumento de ELEV. Como há apenas uma variável preditora na FPESe, não há necessidade de preocupação com a ocorrência de multicolinearidade (Tabela 33).

Tabela 31 – Coeficientes da função de predição espacial da log-razão aditiva ln(silte/areia) no domínio fisiográfico superior (*n* = 150).

ln(silte/areia)	Estimativa	Erro padrão	Valor t	Pr(>|t|)
Intercepto	-3,067589	0,475824	-6,447	1,52e-09
ELEV[1]	0,008448	0,001181	7,152	3,66e-11
Erro quadrático médio: 0,704 com 148 graus de liberdade				
R^2 múltiplo: 0,2568; R^2 ajustado: 0,2518				
Critério de Informação de Akaike: -103,3237				

[1] Variáveis preditoras: ELEV – elevação.

Tabela 32 – Análise de variância da função de predição espacial da log-razão aditiva ln(silte/areia) no domínio fisiográfico superior (*n* = 150).

Fonte de variação	Graus de liberdade	Soma de quadrados	Quadrado médio	F	Pr(>F)
Regressão[1]	1	25,346	25,3462	51,147	3,663e-11
ELEV	1	25,346	25,3462	51,147	3,663e-11
Resíduos	148	73,343	0,4956	-	-
Total	149	98,689	0.6623	-	-

[1] Variáveis preditoras: ELEV – elevação.

Tabela 33 - Matriz variância-covariância dos parâmetros ajustados da função de predição espacial da log-razão aditiva ln(silte/areia) no domínio fisiográfico superior (*n* = 150) e fator de inflação da variância (FIV) associado a cada variável preditora.

	Intercepto	ELEV	FIV[2]
Intercepto	0,226408714	-5,579380e-04	-
ELEV[1]	-0,000557938	1,395284e-06	-

[1] Variáveis preditoras: ELEV – elevação.
[2] O cálculo do fator de inflação da variância exige a existência de duas ou mais variáveis.

Novamente não há relação significativa entre os res_i e as predições e nenhuma observação foi eliminada do conjunto de dados para o ajuste dos modelos de regressão (Figura 27). Apesar de haverem observações com $res_i > |2{,}0|$, $res_i^* > 1{,}96$ e $h_{ii} > 3p/n = 3\times4/150 = 0{,}08$, não houveram valores de D_i superiores a 1,0.

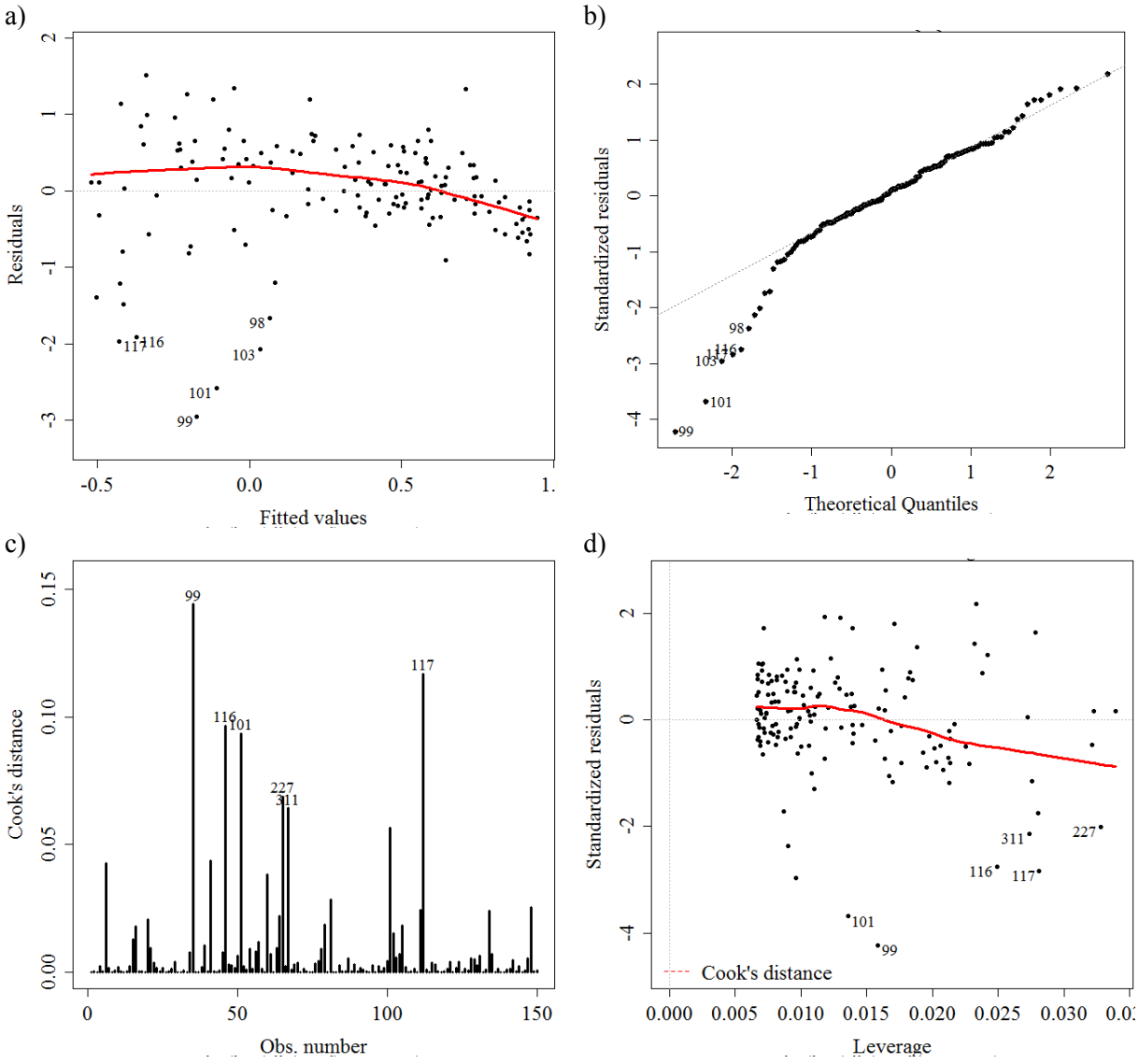


Figura 27 – Análise gráfica da qualidade do ajuste da função de predição espacial da log-razão aditiva ln(silte/areia) no domínio fisiográfico superior (n = 150): a) valores preditos (*fitted values*) e resíduos (*residuals*), b) quantis teóricos (*theoretical quantiles*) e resíduos padronizados (*standardized residuals*), c) distância de Cook (*Cook's distance*) associada a cada observação (*obs. number*), d) alavancagem (*leverage*) e resíduos padronizados (*standardized residuals*).

Dentre as seis observações que apresentaram os maiores res_i, cinco são provenientes de solos derivados de rochas e materiais sedimentares (observações nº 99, 101, 103, 116 e 117) e apenas uma é proveniente de solo derivado de rocha ígnea extrusiva (observação nº 98) (Tabela 34). Quatro delas já também apresentaram res_i elevados em outras FPESe descritas acima (observações nº 99, 101, 116 e 117). Novamente as observações provenientes de solos derivados de rochas e materiais sedimentares apresentaram os res_i mais significativos. No que diz respeito à observação nº 98, o teor de argila é bastante reduzido em comparação às demais amostras localizadas na mesma elevação.

Tabela 34 – Identificação e característics das observações atípicas e influenciais na função de predição espacial da log-razão aditiva ln(silte/areia) no domínio fisiográfico superior.

ID	Material de origem	Classe de solo	Areia (%)	Silte (%)	Argila (%)	CONV	ELEV	IPE
98	Ígnea	RL	69	14	17	-4,47202	370,9979	3,397063
99	Sedimentar	RQ	91	4	5	-3,9302	342,751	6,123335
101	Sedimentar	RL	88	6	6	-0,53634	350,3945	5,692258
116	Sedimentar	PV	88	9	3	0,518997	319,407	6,30546
227	Sedimentar	RL	80	12	8	3,605218	303,6161	6,790433
311	Sedimentar	RL	80	12	8	-1,04097	314,1833	6,978386

Nota: ID – identificação da observação, RL – Neossolo Litólico, RQ – Neossolo Quartzarênico, PV – Argissolo Vermelho, CONV – índice de convergência (%), ELEV – elevação (m), IPE – logaritmo natural do índice de potência de escoamento (adimensional).

Com relação às seis observações mais influentes, todas elas (observações nº 99, 101, 116, 117, 227 e 311) são provenientes de solos derivados de rochas e materiais sedimentares, todas localizadas bastante próximo da cota de 300 m, estabelecida como limítrofe entre os domínios fisiográficos estudados.

As estatísticas da validação cruzada (174 observações, 10 partições, 100 repetições) mostram que as FPESe construídas para estimar as log-razões aditivas ln(argila/areia) e ln(silte/areia) explicam entre 11% e 40% da variância da distribuição do tamanho de partículas do solo (Tabela **35**). As melhores predições são as da fração argila e areia, para a qual a FPESe é capaz de explicar, respectivamente, 40% (R^2_{aj} = 0,40) e 36% (R^2_{aj} = 0,36) da variância. As predições mais pobres são agora aquelas da fração silte (R^2_{aj} = 0,11). Note que os erros de predição (*RMSE*) são novamente maiores para a fração areia, mas quando

avaliados proporcionalmente ($RMSE_{norm}$), todos ficam entre 0,14 e 0,17. Quanto aos erros médios (*EM*) de predição, as frações silte e areia são sobreestimadas (*EM* = 0,36% e *EM* = 0,10%), enquanto a fração argila é subestimada (*EM* = -0,46%).

Tabela 35 – Estatísticas da validação cruzada (174 observações; 10 partições; 100 repetições) das funções de predição espacial da distribuição do tamanho de partículas no domínio fisiográfico superior.

Fração de tamanho de partícula	Estatísticas[1]	Percentil 2,5	Mediana	Percentil 97,5
Areia	*RMSE* (%)	13.4261264	13.5255906	13.6720396
	$RMSE_{norm}$	0.1579544	0.1591246	0.1608475
	EM (%)	0.2885150	0.3565719	0.4156943
	R^2_{aj}	0.3547289	0.3604753	0.3665951
Silte	*RMSE* (%)	11.23422485	11.3289992	11.5146443
	$RMSE_{norm}$	0.17553476	0.1770156	0.1799163
	EM (%)	0.02398648	0.1003501	0.1932704
	R^2_{aj}	0.11067350	0.1153701	0.1241767
Argila	*RMSE* (%)	7.5576532	7.6252677	7.7451340
	$RMSE_{norm}$	0.1481893	0.1495151	0.1518654
	EM (%)	-0.5261717	-0.4587483	-0.3847395
	R^2_{aj}	0.3905845	0.4016966	0.4162169

[1] *RMSE* – raiz quadrada do erro quadrático médio, $RMSE_{norm}$ – raiz quadrado do erro quadrático médio normalizada, *EM* – erro médio, R^2_{aj} - coeficiente de determinação ajustado.

A Figura 28 mostra os valores preditos plotados contra os valores medidos das três frações de tamanho de partícula. A dispersão dos pontos ao redor da linha 1:1 é acentuada, mas seguindo a sua tendência, ao contrário das predições feitas no domínio fisiográfico inferior, onde os pontos aparecem deslocados para a direita (silte e argila) ou para a esquerda (areia). Notadamente a fração argila é a melhor predita, confirmando as estatísticas da validação cruzada (Tabela 35).

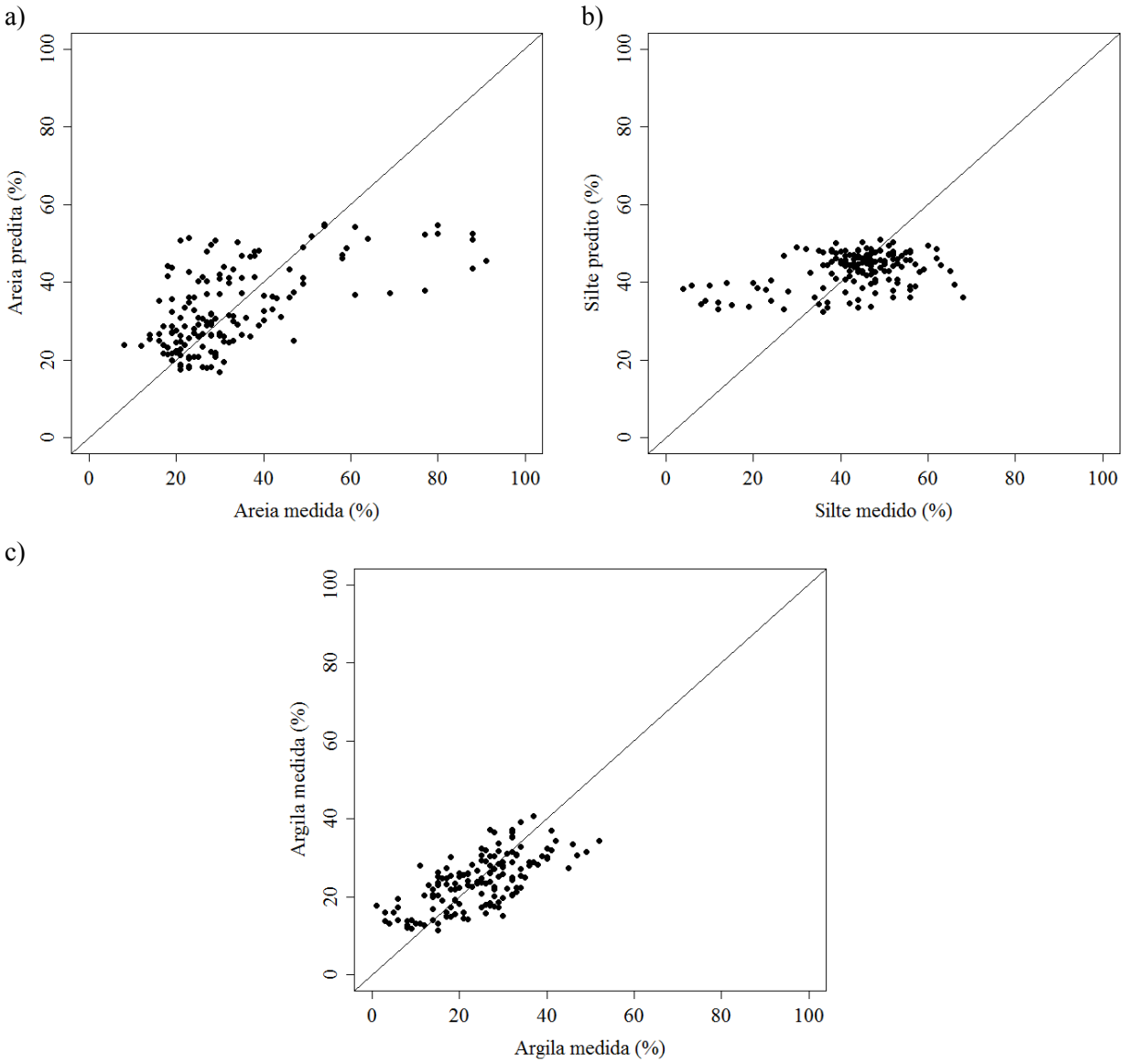


Figura 28 – Distribuição do tamanho de partícula (a – areia, b – silte, c – argila) medida e predita pelas funções de predição espacial da distribuição do tamanho de partículas do solo no domínio fisiográfico superior (n = 150).

6.2.3 Predições e seus resíduos

Ao avaliar, em cada domínio fisiográfico, as FPESe construídas para estimar as log-razões aditivas em toda a área de estudo, verifiquei que seu desempenho foi superior ao das FPESe construídas para cada domínio fisiográfico. No domínio fisiográfico inferior as FPESe explicaram entre 14% e 14% da variância (Tabela 36), enquanto que as FPESe construídas especificamente para esse domínio explicaram apenas de 8% a 12% da variância da

distribuição do tamanho de partículas do solo (Tabela 26). Os maiores erros de predição (*RMSE*) foram os da fração areia, mas proporcionalmente todos ficaram ao redor de 20% do intervalo de variação dos dados ($RMSE_{norm}$ = 0,20). Note que no domínio inferior a fração areia é sobreestimada (*EM* > 0), enquanto as frações silte e argila são subestimadas (*EM* < 0).

Tabela 36 – Estatísticas da validação cruzada (10 partições; 100 repetições) das funções de predição espacial da distribuição do tamanho de partículas em toda a área de estudo aplicadas aos domínios fisiográficos inferior (*n* = 165) e superior (*n* = 174).

Fração de tamanho de partícula	Estatísticas[1]	---------- Inferior ----------			---------- Superior ----------		
		Percentil 2,5	Mediana	Percentil 97,5	Percentil 2,5	Mediana	Percentil 97,5
Areia	*RMSE* (%)	14,758389	14,844407	14,981148	13,923409	13,986467	14,056771
	$RMSE_{norm}$	0,1941893	0,1953211	0,1971204	0,1638048	0,1645467	0,1653738
	EM (%)	1,3355863	1,3876051	1,4464821	1,4359085	1,4839867	1,5510953
	R^2_{aj}	0,1549777	0,1588235	0,1646179	0,4970299	0,5025849	0,5085174
Silte	*RMSE* (%)	10,390740	10,468327	10,588348	12,451235	12,553104	12,663123
	$RMSE_{norm}$	0,1855489	0,1869344	0,1890776	0,1945506	0,1961423	0,1978613
	EM (%)	-1,013076	-0,957416	-0,901008	-1,139352	-1,064363	-0,983702
	R^2_{aj}	0,1343999	0,1405048	0,1502547	0,3786567	0,3910683	0,4060800
Argila	*RMSE* (%)	5,3421255	5,3863732	5,4784475	7,8966818	7,9718168	8,0615650
	$RMSE_{norm}$	0,1907902	0,1923705	0,1956588	0,1548369	0,1563101	0,1580699
	EM (%)	-0,478829	-0,428996	-0,391714	-0,494818	-0,418658	-0,351255
	R^2_{aj}	0,1650630	0,1734171	0,1849666	0,1677756	0,1749701	0,1843328

[1] *RMSE* – raiz quadrada do erro quadrático médio, $RMSE_{norm}$ – raiz quadrado do erro quadrático médio normalizada, *EM* – erro médio, R^2_{aj} - coeficiente de determinação ajustado.

Já no domínio fisiográfico superior, as FPESe construídas para toda a área de estudo explicaram entre 17% e 50% da variância (Tabela 36), enquanto as FPESe construídas especificamente para esse domínio fisiográfico explicaram entre 11% e 40% da variância da distribuição do tamanho de partículas do solo (Tabela 35). Os maiores erros de predição (*RMSE*) são, novamente, os da fração areia, enquanto a argila é a que possui os menores *RMSE*. Ao contrário do que ocorre no domínio fisiográfico inferior, as frações silte e argila são sobreestimadas e a fração areia é subestimada.

Ao analisar as anteriores que descrevem as observações atípicas e influenciais (Tabela

12, Tabela 16, Tabela 21, Tabela 25 e Tabela 30), é possível observar que as observações que apresentaram os maiores res_i relacionadas às predições da fração silte foram aquelas de nº 99, 101 e 135, enquanto as mais influenciais foram aquelas de nº 1, 162 e 302. Todas elas são provenientes de solos Neossolos que, quando derivados de rochas ígneas extrusivas possuem ELEV < 300 m e, quando derivados de rochas e materiais sedimentares possuem ELEV > 300 m. A exceção é o solo Neossolo Flúvico, que possui ELEV < 300 m.

Já as observações que apresentam os maiores res_i relacionados às predições da fração argila são aquelas de nº 99, 116, 121, 162 e 328, enquanto as mais influenciais são aquelas de nº 160, 162, 302 e 328. A maioria delas é proveniente de solos Neossolos Litólicos derivados de rochas ígneas extrusivas com ELEV < 300 m. As demais são provenientes de solos derivados de rochas e materiais sedimentares com ELEV > 300 m.

A análise geoestatística dos res_i da predição da distribuição do tamanho de partículas do solo em toda a área de estudo mostra que existe dependência espacial (Figura 29). Os pares de pontos separados por distâncias mais curtas possuem menor semivariância (são mais parecidos), enquanto pares de pontos separados por distâncias maiores apresentam maior semivariância (são menos parecidos). Além disso, o alcance dos semivariogramas experimentais ajustados (areia = 888 m; silte = 1493 m; argila = 2977 m) é significativamente superior a distância entre os pontos amostrados, para os quais a distância média mínima de separação é de 181 m, variando de 18 a 328 m. Por fim, o Grau de Dependência Espacial (*GDE*) dos res_i das três frações de tamanho de partícula é classificado como moderado, uma vez que os valores estão na faixa entre 25% e 75% (areia = 49,87%; silte = 65,84%; argila = 45,20%), indicando que sua interpolação via krigagem é possível (Tabela 37).

Tabela 37 – Estatísticas da análise variográfica (10 lags de 300 m) dos resíduos de predição da distribuição do tamanho de partículas do solo em toda a área de estudo (n = 339).

	Areia	Silte	Argila
Modelo experimental	Exponencial	Exponencial	Exponencial
Efeito pepita	57,426	45,847	26,791
Alcance (m)	888,109	1493,119	2977,414
Contribuição	57,135	88,380	22,093
Patamar	114,561	134,227	48,884
GDE (%)	49,87	65,84	45,19

Nota: GDE – Grau de Dependência Espacial.

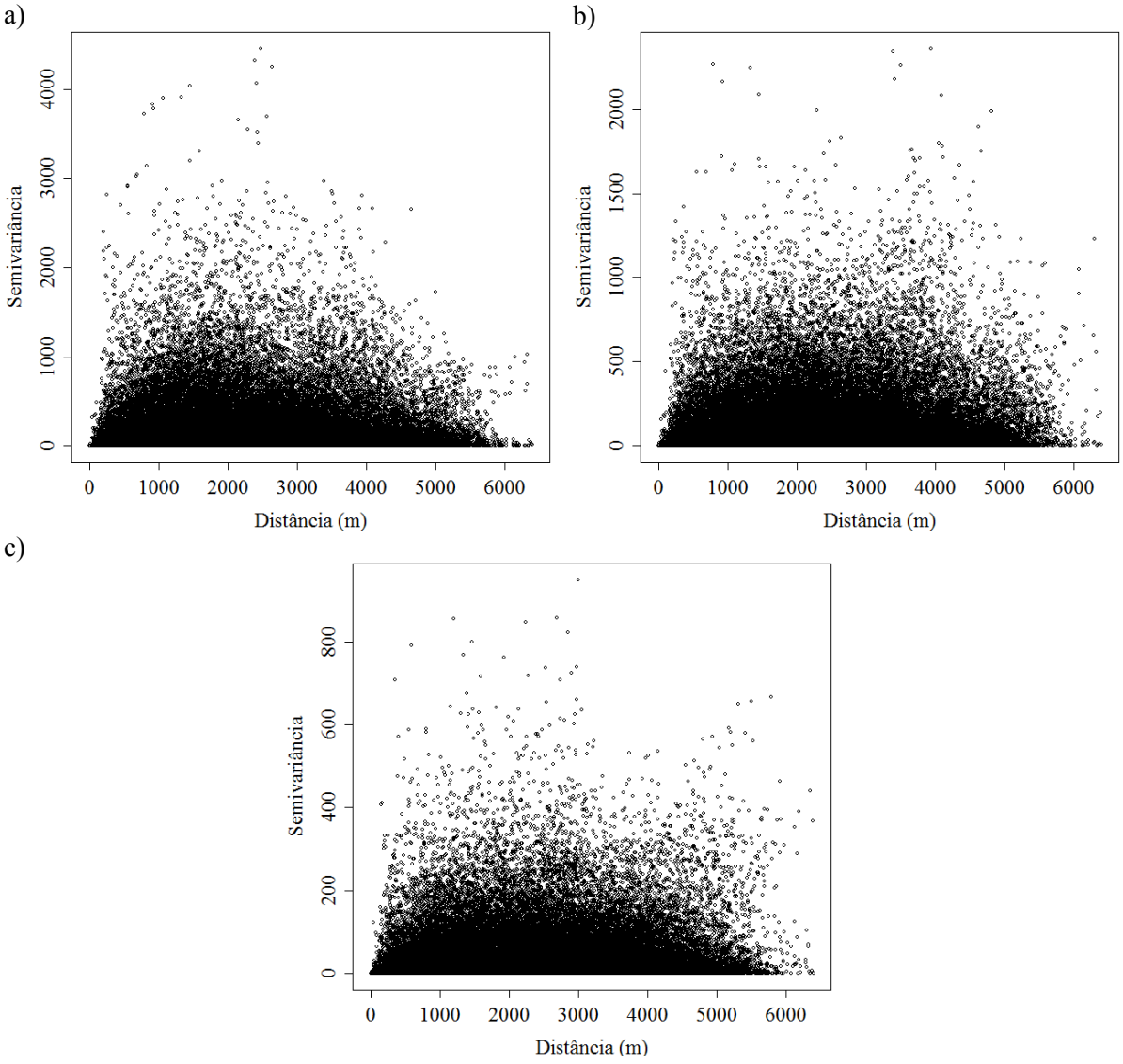


Figura 29 – Nuvem de variograma dos resíduos da predição da distribuição do tamanho de partícula em toda a área de estudo.

A Figura 30 mostra que os res_i das predições da distribuição do tamanho de partículas feitas pelas FPESe construídas para toda a área de estudo são mais significativos em três regiões. A primeira delas é a região Norte-Nordeste da área de estudo, onde ocorrem as área de maior ELEV (Figura 16) e solos derivados de rochas ígneas extrusivas, predominantemente Argissolos Vermelho-Amarelos (Miguel, 2010). Trata-se de uma área cujas características de relevo, geologia e solos se assemelham bastante àquelas da região fisiográfica adjacente localizada ao Norte (o Planalto). Como essa porção se trata da mais elevada da paisagem (Figura 16), e a relação entre a proporção da fração areia e a variável preditora ELEV (que explica a maior parte da variância) é inversa (Tabela **9** e Tabela **13**), as

predições feitas pelas FPESe geram subestimativas. Já no caso das frações silte e argila, por terem uma relação positiva com a variável preditora ELEV, as predições feitas pelas FPESe geram sobreestimativas.

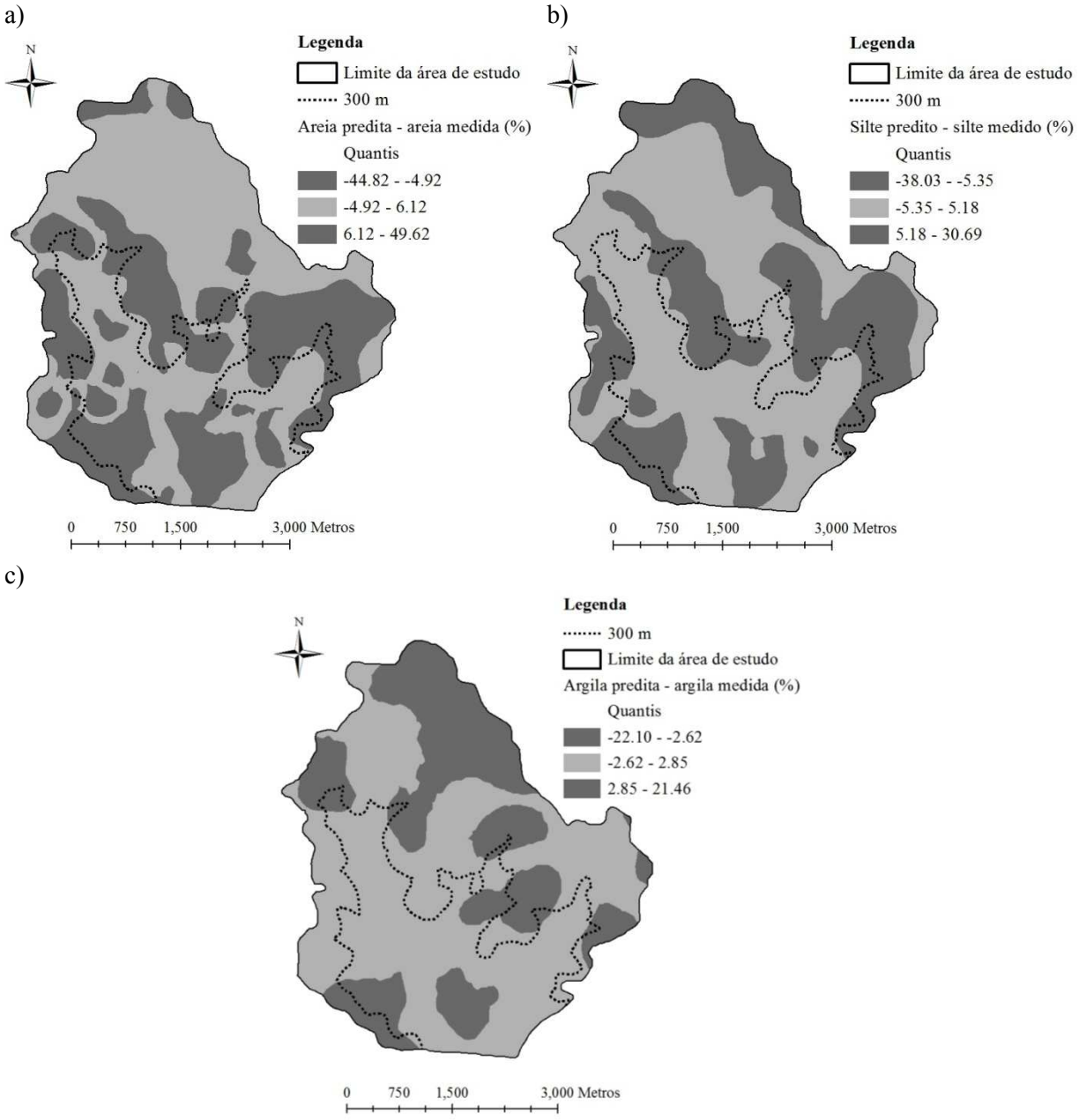


Figura 30 – Resíduos krigados da predição da distribuição do tamanho de partículas (a – areia, b – silte, c – argila) utilizando as FPESe construídas para toda a área de estudo.

A segunda região onde os res_i são maiores compreende o Sul da área de estudo, próximo do exudório. Esse setor apresenta os menores valores de ELEV e maiores de IUT

(Figura 16), com solos derivados apenas de rochas e materiais sedimentares (colúvios e depósitos fluviais recentes), predominantemente Planossolos Háplicos e Neossolos Flúvicos (Miguel, 2010). Como essa se trata da porção mais baixa da paisagem (Figura 16), e a relação entre a proporção da fração areia e a variável preditora ELEV é inversa (Tabela 9 e Tabela 13), as predições feitas pelas FPESe geram sobreestimativas. Já no caso das frações silte e argila, por terem uma relação positiva com a variável preditora ELEV, as predições feitas pelas FPESe geram subestimativas.

A terceira e última região em que ocorrem os maiores res_i é aquela que compreende toda a área ao redor da cota de 300 m (linha pontilhada). Trata-se de uma região em que ocorrem solos derivados de diversos materiais de origem, sejam eles de origem sedimentar ou ígnea, cuja expressão é descontínua, levando a diferenças marcantes nas características dos solos formados. Além disso, essas áreas são bastante acidentadas, possuindo elevados valores de IRS (Figura 16). Conforme visto na análise dos r_i nas seções acima, nos locais imediatamente acima da cota de 300 m são as observações provenientes de solos derivados de rochas e materiais sedimentares (derivados da Formação Botucatu) aquelas que apresentam os maiores r_i. Nesse caso, a proporção da fração areia é subestimada em favor das demais. Já nos locais imediatamente abaixo da cota de 300 m são as observações provenientes de solos (Neossolos Litólicos) derivados de rochas ígneas extrusivas (derivados da Sequência Inferior da Formação Serra Geral) aquelas que apresentam os maiores res_i. Nesse caso, as frações silte e argila são subestimadas em favor da fração areia.

7 DISCUSSÃO

7.1 Predição da distribuição do tamanho de partículas do solo

Em trabalho recente, ten Caten et al. (2011d) mostraram que é possível construir FPESe para o mapeamento preditivo de classes de solo do Sistema Brasileiro de Classificação de Solos na região do Rebordo do Planalto do RS. Os resultados que encontrei mostram que também é possível construir FPESe para estimar a distribuição do tamanho de partículas dos solos dessa região utilizando atributos de terreno como variáveis preditoras. Isso significa que o MDE que utilizei deve representar satisfatoriamente a paisagem da área de estudo, uma vez que os atributos de terreno dele derivados foram capazes de capturar a relação entre os solos e o ambiente em que ocorrem.

Ao analisar as FPESe que construí é possível perceber que elas refletem o efeito de dois importantes fatores de formação do solo: o material de origem e o relevo. O efeito do material de origem é reproduzido por ELEV. Isso porque a distribuição das formações geológicas da área de estudo (e da região do Rebordo do Planalto do RS como um todo) se dá na forma de "camadas" sobrepostas (Sartori, 2009). Conforme mencionei na seção Material e Métodos, em elevações superiores a ±350 m ocorre a Sequência Superior da Formação Serra Geral, em elevações entre ±200 m e ±350 m a Sequência Inferior da Formação Serra Geral e no seu interior ou abaixo dela a Formação Botucatu, enquanto que em elevações abaixo de ±200 m ocorrem a Formação Caturrita, além dos depósitos coluviais e fluviais recentes. Os solos derivados de rochas ígneas extrusivas da Formação Serra Geral possuem distribuição do tamanho de partículas mais fina (Miguel, 2010). Já os solos derivados de rochas e materiais sedimentares possuem distribuição do tamanho de partículas mais grosseira, com predomínio claro da fração areia sobre as outras duas (Miguel, 2010). Assim, os solos das partes mais elevadas da área de estudo possuem distribuição do tamanho de partículas mais fina, enquanto aqueles das partes mais baixas possuem distribuição do tamanho de partículas mais grossa. Como ELEV é a variável preditora que melhor explica essa variação na geologia, tendo explicado a maior parte da variância das FPESe, o feito do material de origem na distribuição do tamanho de partículas dos solos é mais importante do que o efeito do relevo.

O efeito do relevo sobre a distribuição do tamanho de partículas é descrito por CONV e IPE, além da própria ELEV, uma vez que relevo e geologia são intimamente relacionados. Mas se considerarmos que ELEV está mais relacionada ao material de origem do que ao relevo, é possível dizer que o efeito do relevo sobre a distribuição do tamanho de partículas é pequeno, uma vez que CONV e IPE não explicaram juntas mais do que 10% da variância (estimada pela soma de quadrados total) em qualquer das FPESe construídas. Mesmo assim, é possível inferir que as frações mais grosseiras (silte e principalmente areia) tendem a ficar depositadas nas porções mais baixas do terreno, ou seja, nos locais de convergência dos fluxos hídricos (relação positiva com CONV). Além disso, as frações mais grosseiras tendem a predominar nos locais de maior potencial de perda de solo (relação positiva com IPE). Como o IPE está mais relacionado à ocorrência de erosão em canais (Moore et al., 1993), o fluxo hídrico nesses locais proporciona a perda das frações mais finas e o acúmulo das frações mais grosseiras no fundo. Note que IPE é proporcional a DECL e AC, sugerindo que a erosão pode ter sido importante na formação dos solos, removendo as partículas mais finas das superfícies mais declivosas e que recebem maior quantidade de fluxo hídrico. Contudo, não devemos esquecer que o efeito da erosão em sulco e entre sulcos atualmente deve ser pequeno, uma vez que a maior parte da área de estudo possui cobertura vegetal que impede esse tipo de erosão (pastagem e floresta) (Samuel-Rosa et al., 2011b). Contudo, nas décadas anteriores, a erosão do solo teve papel importante na degradação dos solos, havendo inclusive hipóteses de que tenha havido um retrocesso pedogenético de vários solos (Streck et al., 2008; Samuel-Rosa et al., 2011b). Segundo essa hipótese, as perdas de solo foram tamanhas que horizontes inteiros podem ter sido perdidos, levando a classificação dos solos em classes pedogeneticamente menos desenvolvidas. Hoje, uma parte significativa desse material erodido está depositada no fundo do reservatório localizado a jusante, o qual já perdeu 29% da sua capacidade de estoque de água em 29 anos de operação (Dill et al., 2004). Através de um novo levantamento batimétrico do reservatório poderia ser verificado se a revegetação que vem ocorrendo na área de estudo nos últimos anos teve efeito sobre a redução dos processos erosivos.

Mas se o efeito do relevo sobre a distribuição do tamanho de partículas do solo é significativamente inferior àquele do material de origem, ainda pode-se sugerir que sua expressão é diferente em cada um dos domínios fisiográficos. Tomemos de início as FPESe construídas para toda a área de estudo, onde a sequência de importância das variáveis preditoras é a seguinte: ELEV > IPE > CONV. A importância de ELEV como atributo de terreno é que está diretamente relacionado à suscetibilidade dos solos aos processos erosivos,

uma vez que é um indicativo da energia potencial de uma superfície (Wilson e Gallant, 2000a). Já IPE nada mais é do que um índice da capacidade de transporte de sedimento em uma dada superfície, enquanto CONV está relacionado à direção dos fluxos hídricos e, portanto, às zonas de perda e acumulação (Moore et al., 1993). Sendo a sequência de importância das variáveis preditoras ELEV > IPE > CONV, o principal efeito do relevo sobre a distribuição do tamanho de partículas é através do controle da erosão, com maior peso para os processos de perda e transferência das partes mais elevadas (ELEV e IPE), seguido do processo de acumulação nas partes mais baixas (CONV). No domínio fisiográfico superior a sequência de importância das variáveis preditoras é similar (ELEV > IPE), exceto pelo fato de que CONV não aparece em nenhuma FPESe. Isso indica que os processos de perda e transferência são os que prevalecem e que as zonas de acumulação são pouco significativas. O contrário ocorre no domínio fisiográfico inferior, onde CONV possui maior importância nas FPESe construídas (ln(argila/areia): ELEV > CONV > IPE; ln(silte/areia): CONV > IPE > ELEV). Isso indica que o controle do relevo sobre a distribuição do tamanho de partículas do solo nesse domínio fisiográfico se dá, sobretudo, pela constituição de zonas de acumulação. Já os processos de perda e transferência são pouco expressivos. Apesar dessas relações, ressalto que esse efeito (do relevo e do processo erosivo) é pequeno se comparado ao efeito do material de origem.

De qualquer maneira, a concretização da possibilidade de construir FPESe na região do Rebordo do Planalto do RS se deve, sobretudo, pelo fato da região ser heterogênea em relação aos atributos do terreno. Em locais onde há uma forte homogeneidade do relevo é mais difícil construir FPESe (Sumfleth e Duttmann, 2008). Entretanto, devido à elevada variabilidade natural da distribuição do tamanho de partículas e aos erros associados a ambos a determinação laboratorial da distribuição do tamanho de partículas dos solos e a coleta dos atributos do terreno (sobretudo devido aos erros presentes no MDE utilizado), não é de se esperar que FPESe possam explicar mais de aproximadamente 70% da variância (Moore et al., 1993). Moore et al. (1993) realizou um dos primeiros trabalhos de mapeamento preditivo de propriedades do solo utilizando atributos de terreno como preditores em uma toposequência de solos geologicamente homogênea. Mesmo assim suas FPESe não conseguiram explicar mais de 52% e 64% da variância das frações de tamanho areia e silte, respectivamente. Mesmo os mapas tradicionais de solos (baseados em classes de solos com limites abruptos) geralmente explicam aproximadamente metade da variância de propriedades do solo como a distribuição do tamanho de partículas em uma região (Webster e Oliver, 1990). Assim, o desempenho das FPESe que construí para predizer a distribuição do tamanho

de partículas do solo em toda a área de estudo pode ser considerado satisfatório, uma vez que explicam ao redor da metade da variância. Isso comprova minha primeira hipótese. Entretanto, é preciso destacar que as estatísticas da validação cruzada podem conter erros, uma vez que os pontos amostrais não constituem uma amostra aleatória da população (Brus et al., 2011). Uma amostragem probabilística adicional deveria ser realizada para fornecer estimativas mais acuradas do erro de predição.

No que diz respeito às variáveis preditoras que utilizei para construir as FPESe (CONV, ELEV, IPE e IUT), elas estão entre as mais utilizadas em estudos de mapeamento preditivo de propriedades do solo (McBratney et al., 2003; Bishop e Minasny, 2006). A eliminação de 73% das variáveis preditoras do processo analítico antes mesmo de iniciar o ajuste dos modelos de regressão mostra a clara existência de multicolinearidade entre os diversos atributos de terreno derivados do MDE, sobretudo entre primários e secundários. Isso porque os últimos são obtidos a partir dos primeiros e assim acabam sendo redundantes por carregarem informações semelhantes (Hengl e MacMillan, 2009), como é o caso de DECL, AC e IUT (Equação (14)). Além disso, atributos de terreno que descrevem características similares do terreno (como PLAN, PERF, CURV e CONV) também são correlacionados (mesmo que de maneira teórica), indicando que seu uso concomitante no ajuste de um modelo de regressão linear múltipla é desnecessário, podendo ser até mesmo inadequado.

Como as variáveis preditoras CONV, ELEV, IPE e IUT não se mostraram relacionadas, mas sim representativas de grupos de variáveis preditoras que explicam características e/ou processos específicos que ocorrem na paisagem, elas é que foram as escolhidas para a construção das FPESe (note que nessas variáveis preditoras estão diretamente incluídas outros atributos primários de terreno, como DECL, AC, PLAN, PERF, entre outros). Isso indica que as variáveis preditoras utilizadas são efetivamente independentes, um dos pressupostos básicos da análise de regressão linear múltipla (Hair et al., 2010). Contudo, é comum encontrar trabalhos de mapeamento preditivo que utilizam como variáveis preditoras atributos de terreno que costumam estar correlacionados (ao menos de maneira teórica) para ajustar modelos de regressão linear múltipla. Os mais comuns são IUT, AC e DECL. Como exemplos podem ser citados os trabalhos de Moore et al. (1993), McKenzie e Ryan (1999), Park e Vlek (2002), Sumfleth e Duttmann (2008), Mendonça-Santos et al. (2010), ten Caten et al. (2011d), entre muitos outros. A maioria desses trabalhos não apresenta estatísticas que permitam avaliar, com segurança, a ocorrência de multicolinearidade, como é o caso da matriz variância-covariância dos coeficientes dos parâmetros ajustados do modelo de regressão e o fator de inflação da variância. Isso indica

que parte das FPESe construídas pode apresentar sérios problemas de multicolinearidade, o que pode comprometer seriamente a sua qualidade (Hair et al., 2010). Cabe verificar se o prejuízo da ocorrência de multicolinearidade nessas FPESe afeta a qualidade das predições realizadas (e em que extensão) e, principalmente, a interpretação ambiental e pedológica de seus resultados. Isso é importante porque as FPESe devem, mais do que resultar em predições acuradas, contribuir para a construção do conhecimento pedológico (Milne e Lark, 2008).

7.2 Fatores afetando o desempenho das FPESe

Devido à elevada complexidade geológica da região do Rebordo do Planalto do RS, esperava que a estratificação em dois domínios fisiográficos mais homogêneos do ponto de vista geológico pudesse resultar na construção de FPESe com maior capacidade preditiva, uma vez que esse é o procedimento recomendado pela literatura (Gessler et al., 1995). Conforme enunciado em minha segunda hipótese, essa estratificação deveria levar a um desempenho de até aproximadamente 70%. Contudo, os resultados não atenderam às minhas expectativas, negando minha segunda hipótese. De maneira geral, as FPESe que construí para estimar a distribuição do tamanho de partículas em toda a área de estudo (n = 300) tem desempenho superior do que aquelas que construí separadamente para cada domínio fisiográfico (n = 150). Uma das razões do desempenho inferior das FPESe que construí individualmente para cada domínio fisiográfico deve ser a maior homogeneidade dos atributos do terreno resultante da estratificação (Sumfleth e Duttmann, 2008) (apesar de não ter realizado qualquer avaliação da variação dos atributos de terreno nos dois domínios fisiográficos). Entretanto, o desempenho das FPESe foi diferente em cada domínio fisiográfico. Todas as predições foram mais acuradas no domínio fisiográfico superior, seja usando as FPESe que construí utilizando os dados de toda a área de estudo (n = 300) ou as FPESe que construí utilizando os dados de cada domínio fisiográfico (n = 150). Isso sugere que a maior homogeneidade dos atributos do terreno não é o principal fator afetando a capacidade preditiva das FPESe.

O fator que acredito definir em maior grau o desempenho das FPESe que construí é complexidade geológica. Assim sendo, o melhor desempenho das FPESe no domínio fisiográfico superior seria devido a maior homogeneidade geológica ali encontrada. Rochas ígneas extrusivas da Sequência Superior da Formação Serra Geral predominam no domínio

fisiográfico superior, onde apenas pequenas manchas de solos derivados da Formação Botucatu são encontradas no seu limite inferior (próximo da cota de 300 m). Além disso, o domínio fisiográfico superior constitui uma superfície geomórfica mais estável. Em tais condições a relação entre a distribuição do tamanho de partículas e os atributos de terreno parece ser mais evidente, permitindo a construção de FPESe com desempenho superior.

Por outro lado, o domínio fisiográfico inferior constitui uma superfície geomórfica mais jovem, com diversas áreas deposicionais (zonas de acumulação) e que possui maior heterogeneidade geológica. Apesar da maioria dos materiais de origem encontrada nesse domínio fisiográfico ser de origem sedimentar, eles são de cinco tipos: arenitos eólicos (Formação Botucatu), arenitos fluviais (Formação Caturrita), colúvios das Formações Serra Geral e Botucatu, colúvios das Formações Botucatu e Caturrita, e depósitos fluviais recentes. Os solos desenvolvidos a partir dos colúvios e depósitos fluviais recentes são bem mais jovens do que aqueles formados nas condições mais estáveis do domínio fisiográfico superior. Note que uma das observações que apresentou os maiores res_i e influenciou os modelos de regressão ajustados é aquela de nº 1, proveniente de solo Neossolo Flúvico. Além disso, os solos desenvolvidos a partir dos arenitos localizados nos terços médio e superior das coxilhas foram fortemente alterados pela erosão causada pelas práticas agrícolas inadequadas usadas na região durante várias décadas (Samuel-Rosa et al., 2011b). Como consequência desses fatores a distribuição do tamanho de partículas do solo possui uma relação moderada com a maioria dos atributos de terreno. Não é de se esperar que o desempenho de FPESe construídas em tais condições seja melhor do que aquele observado no presente estudo.

O efeito negativo dessa complexidade geológica também se expressa na região de transição entre os domínios fisiográficos em estudo (ao redor da cota de 300 m). Lembre que as observações que se apresentaram mais atípicas e influenciais no domínio fisiográfico superior são aquelas provenientes de solos derivados de rochas e materiais sedimentares, ao passo que no domínio fisiográfico inferior as observações mais atípicas e influenciais são aquelas provenientes de solos derivados de rochas ígneas extrusivas. Em primeiro lugar, a ocorrência desses solos é descontínua, irregular, nunca paralela ao plano horizontal (conforme ocorre com as "camadas" dos diferentes materiais de origem). Assim, apesar da variável preditora ELEV descrever o efeito do material de origem nas FPESe, ela é incapaz de descrever de maneira acurada toda essa variação dos solos. Além disso, toda a superfície ao redor da cota de 300 m possui elevada rugosidade (Figura 16), sugerindo que sua estabilidade deve ser reduzida, o que dificulta encontrar relações evidentes entre atributos do terreno e a distribuição do tamanho de partículas do solo.

Em segundo lugar, parte significativa das observações mais atípicas (com maiores res_i) e influenciais é proveniente de solos Neossolos Litólicos, derivados de rochas ígneas extrusivas, localizados em ELEV entre 200 e 300 m. Essa faixa de ELEV corresponde àquela de ocorrência da Sequência Inferior da Formação Serra Geral, caracterizada por rochas básicas (basalto), em contraposição às rochas ácidas da Sequência Superior (riolito). Essas duas rochas possuem composição química diferenciada, sendo a primeira relativamente mais pobre em sílica (SiO4 < 52%) do que a segunda (SiO4 > 65%), o que possui influência direta sobre suas características físicas (Pedron, 2007). Uma delas é a resistência ao intemperismo e, portanto, a distribuição do tamanho de partículas do solo formado. Nesse caso, solos derivados de basalto devem ter distribuição do tamanho de partículas mais fina do que solos derivados de riolito, onde a presença de quartzo costuma ser significativa (Pedron, 2007). Assim sendo, temos solos com distribuição do tamanho de partículas mais fina em posições intermediárias da paisagem do que aqueles localizados nas porções de maior ELEV, o que foge a relação positiva encontrada entre proporção da fração de tamanho argila e ELEV nas FPESe construídas. Pesa também o fato de que estamos tratando de Neossolos Litólicos, ou seja, solos pouco desenvolvidos devido à reduzida ação dos processos pedogenéticos, seja pela resistência do material de origem ou pelas condições do clima, relevo e do próprio tempo de atuação dos agentes intempéricos (Pedron, 2007). Como os Neossolos Litólicos em questão localizam-se nas bordas expostas da Sequência Inferior da Formação Serra Geral (o topo está recoberto pelas Formações Botucatu e Sequência Superior da Formação Serra Geral), a ação dos agentes intempéricos torna-se mais dificultada e heterogênea, levando a formação de solos com distribuição do tamanho de partículas variável. Tal variação dificilmente será capturada pelos atributos de terreno que utilizei como variáveis preditoras.

Fora a complexidade geológica, existem outros fatores que devem estar afetando negativamente o desempenho das FPESe construídas. Um deles é o fato de que as regiões Norte-Nordeste e Sul, que apresentaram res_i elevados, são mais similares em termos de características fisiográficas, aos domínios fisiográficos adjacentes (o Planalto ao Norte e a Depressão Central ao Sul) do que à própria região do Rebordo do Planalto. Trata-se das regiões onde ocorrem os valores mais extremos de ELEV (valores mínimos e máximos). Isso sugere que a relação entre ELEV e a distribuição do tamanho de partículas não pode ser descrita pelo mesmo modelo linear em toda a área de estudo. Em ambas as regiões (Norte-Nordeste e Sul) a curva descrevendo a relação entre ELEV e a distribuição do tamanho de partículas é menos inclinada, ou seja, a distribuição do tamanho de partículas varia menos

com a variação de ELEV. Em um cenário mais complexo, é possível dizer que modelos lineares não são os mais adequados para descrever essa relação.

Outro fator que pode ser influencial está relacionado aos conjuntos de dados utilizados para construir as FPESe. E aqui me refiro tando ao procedimento de coleta das amostras de solo adotado Miguel (2010) e Samuel-Rosa el al. (2011b), como ao procedimento que adotei para obter subamostras aleatórias de $n = 300$ e $n = 150$ observações dos conjuntos originais de observações ($n = 339$, $n = 165$, $n = 174$). No que diz respeito ao procedimento de coleta adotado por Miguel (2010) e Samuel-Rosa et al. (2011b), ao ser baseado em seu conhecimento tácito, creio que as observações obtidas sejam significativamente enviesadas. Isso porque, apesar de a maior parte do intervalo de variação dos atributos de terreno na área de estudo ter sido amostrada, algumas feições foram subamostradas, indicando que a amostragem não foi representativa. Os locais onde a amostragem foi menos representativa são exatamente aqueles em que o acesso é mais difícil, seja pela declividade acentuada ou presença de acidentes geográficos (áreas de maior DECL, LS, IPE e IRS). Contudo, também houve locais subamostrados onde a declividade não é acentuada e não existem acidentes geográficos. São as áreas caracterizadas pela maior concavidade: PERF, PLAN, CURV, CONV e IPE. Essas áreas são aquelas localizadas nas porções mais baixas da paisagem, ou seja, os locais para onde convergem os fluxos hídricos (zonas de acumulação), sobretudo próximo aos talvegues. De maneira geral, durante os procedimentos amostrais intencionais (*purposive sampling*), baseados no conhecimento tácito do pedólogo, utilizado nos levantamentos de solos tradicionais, locais como esses não costumam ser amostrados. Isso porque as características do solo nesses locais não costumam ser representativas das manchas de solo mapeáveis na paisagem, cujo tamanho (área mínima mapeável) depende da escala do levantamento de solos. Portanto, em locais de topografia acidentada e com vales acentuados (como na área de estudo em questão), procedimentos amostrais intencionais devem levar a obtenção de conjuntos de dados pouco representativos das áreas de convergência dos fluxos hídricos e de acentuada declividade.

No que diz respeito ao procedimento que adotei para obtenção de subamostras aleatórias de $n = 300$ e $n = 150$ observações dos conjuntos originais de observações, a dúvida está na possibilidade de que o uso de diferentes subamostras aleatórias poderia ter gerado resultados significativamente diferentes. Essa possibilidade constitui uma realidade devido ao fato de os conjuntos originais de observações provavelmente serem enviesados. Infelizmente não encontrei na literatura qualquer trabalho semelhante que permitisse a comparação dos resultados e, sobretudo, uma análise mais crítica.

Por fim, o procedimento de transformação dos dados de distribuição do tamanho de partículas em log-razões aditivas também pode ter afetado as predições. Note que no domínio fisiográfico superior a FPESe construída para estimar a log-razão aditiva ln(argila/areia) teve desempenho superior a daquela para estimar a log-razão aditiva ln(silte/areia). Em termos absolutos, qual é a contribuição de cada uma dessas FPESe para o desempenho das predições das frações de tamanho de partícula (areia, silte e argila) individualmente? Mesmo que Aitchison (1982) tenha mostrado a importância de transformar dados composicionais antes da sua análise estatística, é preciso verificar se esse procedimento não prejudica as predições.

7.3 Melhorando as predições

Apesar do desempenho das FPESe que construí poder ser considerado satisfatório, uma vez que explicaram mais de 50% da variância, ainda há necessidade de melhorias, sobretudo pela discrepância no desempenho entre os dois domínios fisiográficos. Existem algumas alternativas para melhorar as predições da distribuição do tamanho de partículas dos solos da região do Rebordo do Planalto do RS. Essas alternativas estão relacionadas à amostragem dos solos, às variáveis preditoras utilizadas e aos modelos estatísticos com os quais construí as FPESe.

Como já comentei na seção anterior, o procedimento amostral adotado por Miguel (2010) e Samuel-Rosa et al. (2011b) deve possuir forte influência sobre a qualidade das FPESe construídas pelo fato de as observações serem enviesadas. De fato, a fase de coleta das amostras continua sendo o maior gargalo de todo o processo de construção de FPESe, conforme já havia sido denunciado por Webster e Oliver (1990). E isso se deve ao fato de que a maior parte das informações pedológicas de que dispomos hoje é fruto de amostragens intencionais (*purposive sampling*) realizadas durante trabalhos de levantamentos de solos convencionais (GlobalSoilMap.net, 2011). Contudo, não podemos simplesmente descartar essas amostras e fazer novas amostragens probabilísticas para atender os pressupostos dos modelos matemáticos utilizados. Tais amostras, mesmo enviesadas, contêm grande quantidade de informação pedológica, além de terem sido obtidas à custa de quantias significativas de recursos financeiros e atuação de número elevado de pedólogos, o que as torna muito valiosas. Como utilizar esses dados da maneira mais adequada possível é o principal desafio, bem como identificar a necessidade de coleta de amostras extras sempre que necessário.

Assim, a melhoria do desempenho das FPESe que construí pode ser alcançado através da coleta de novas amostras de solo nos locais subamostrados por Miguel (2010) e Samuel-Rosa et al (2011b).

No que diz respeito às variáveis preditoras utilizadas, atributos de terreno ainda mais complexos, que descrevam as bases físicas de outros processos (que não identifiquei nesse estudo devido aos procedimentos que adotei) que possam estar influenciando a distribuição do tamanho de partículas do solo, podem ser uma boa alternativa. Böhner e Selige (2006), por exemplo, desenvolveram três atributos de terreno complexos para determinar a profundidade de extratos quartenários na Alemanha. Os atributos de terreno desenvolvidos, que os autores preferem chamar *parâmetros de processos* são: parâmetro de solifluxão (*solifluction parameter*), parâmetro de balanço de massa (*mass balance parameter*) e parâmetro de umidade (*wetness parameter*). Ao utilizar tais parâmetros em FPESe, construídas a partir de modelos de regressão linear múltipla, os autores conseguiram explicar 88% da variância predita. Entretanto, a derivação de atributos de terreno ainda mais complexos do que os atributos de terreno secundários que conhecemos (IUT, IPE, entre outros) depende da disponibilidade de dados da superfície (MDE) obtidos com alta resolução, o que ainda não é possível (ou viável) para a maior parte do território brasileiro. No caso do estudo realizado por Böhner e Selige (2006), os parâmetros de processos foram derivados de um MDE construído com o uso de dados da superfície obtidos usando o sistema de varredura aerotransportado LIDAR (*Light Detection and Ranging → laser scanning*). O MDE construído a partir desses dados apresentou precisão de 99,3%. Isso mostra que é necessário disponibilizar dados de superfície obtidos com elevada resolução para permitir o desenvolvimento de FPESe mais acuradas. Além disso, poderiam ser utilizadas variáveis preditoras que identifiquem as zonas de perda, transferência e acumulação, uma vez que o controle do relevo sobre os processos erosivos parece agir através da constituição dessas zonas. Até mesmo uma estratificação da área de estudo com base nessas zonas poderia ser realizada. Mas se o efeito do relevo na área de estudo (e, consequentemente, em toda a região do Rebordo do Planalto) sobre a distribuição do tamanho de partículas é pequeno quando comparado ao efeito da geologia, o desenvolvimento de atributos de terreno ainda mais complexos pode não resultar em melhorias no desempenho das FPESe construídas. Exceto se estiverem relacionados mais intimamente a geologia local do que ELEV, variável que expressa bem a variação da geologia na área de estudo.

Há também a possibilidade de uso de dados geológicos (ou mesmo a construção de FPESe para cada unidade geológica) como variáveis preditoras, uma vez que o material de

origem constitui o fator de formação do solo com maior influência sobre a distribuição do tamanho de partículas na região do Rebordo do Planalto do RS. É por esse motivo que Gessler et al. (1995) recomendam a estratificação das áreas de trabalho, sempre que possível, de acordo com o material de origem dos solos. Se o uso de informações geológicas pode reduzir significativamente os erros de predição dos estoques de carbono no solo (Heim et al., 2009), provavelmente os efeitos também são positivos sobre a predição da distribuição do tamanho de partículas. Entretanto, os mapas geológicos disponíveis para a região do Rebordo do Planalto do RS são de acurácia limitada, uma vez que os mesmos foram produzidos com base nas informações das cartas planialtimétricas publicadas na escala de 1:25.000 (Maciel Filho, 1990), a partir do qual gerei o MDE e derivei os atributos de terreno. A maioria das áreas de material coluvial e de depósitos fluviais recentes não está identificada nesses mapas. Ao mesmo tempo, a distribuição espacial desses materiais de origem parece ter uma relação pobre com a maioria dos atributos de terreno, o que dificulta a sua separação na paisagem. Consequentemente é necessário gerar informações geológicas acuradas enquanto for objetivo melhorar as predições da distribuição do tamanho de partículas, assim como de outras propriedades do solo.

Mas além de variáveis preditoras relacionadas aos fatores relevo e material de origem (*r* e *p* do modelo *scorpan*), podem ser utilizadas outras que estejam relacionadas aos demais fatores do modelo scorpan, como organismos (mapas de uso da terra), solo (mapas de classes de solos), clima e idade. Kerry e Oliver (2007), por exemplo, utilizaram com sucesso dados extraídos de fotografias aéreas coloridas para auxiliar na definição dos intervalos amostrais para predição do teor de argila do solo via krigagen. Brown (2007) mostrou que a resposta espectral dos solos no infravermelho próximo também são úteis na predição da proporção da fração argila no solo. As FPESe construídas conseguiram explicar entre 64% e 66% da variância. Entretanto, não se deve esperar que qualquer variável preditora possua maior poder explicativo do que aquelas relacionadas à geologia da área de estudo. A exceção pode ser a entrada de dados de mapas de classes de solos, os quais estão diretamente relacionados ao material de origem.

A última possibilidade em termos de variáveis preditoras a serem utilizadas é a construção de FPESe a partir do ajuste de modelos de regressão a componentes principais. Através desse método é possível obter variáveis preditoras que expliquem a maior parte da variância dos dados e que não são correlacionadas entre si (ten caten et al., 2011a). Assim sendo, sua principal vantagem é a garantia de que os modelos de regressão linear ajustados não correm o risco de serem prejudicados pela ocorrência de multicolinearidade. Entretanto,

como a análise de componentes principais nada mais é do que um procedimento matemático de transformação dos dados, as componentes principais não possuem qualquer significado físico, químico ou biológico (Hengl e Rossiter, 2003). Como resultado, mesmo que seja possível construir FPESe com elevada capacidade preditiva, dificilmente se encontrará qualquer significado pedológico ou ambiental nos resultados obtidos. Devido a esse inconveniente, considero que o uso de componentes principais como variáveis preditoras para construção de FPESe é inadequado, uma vez que deve permanecer como objetivo fundamental do pedólogo a obtenção de resultados que permitam a melhor compreensão da atuação do solo como componente ambiental.

No que diz respeito aos modelos matemáticos que podemos utilizar para construir as FPESe, são duas as possibilidades. A primeira delas se refere à construção de FPESe com modelos matemáticos não-lineares e mais complexos como as redes neurais artificiais e as árvores de regressão. Isso porque a relação entre as propriedades do solo e os atributos de terreno geralmente não é perfeitamente linear (Grunwald, 2006). Os resultados que obtive mostraram que a relação entre ELEV e a distribuição do tamanho de partículas varia ao longo da área de estudo. Entretanto, a literatura mostra que o uso de modelos não-lineares e mais complexos não é uma garantia de que qualquer melhoria nas predições será alcançada. Park e Vlek (2002) compararam o uso de redes neurais artificiais, árvores de regressão e regressão linear generalizada de acordo com a sua habilidade em predizer a fração silte em cinco profundidades. Os resultados foram similares e os modelos de regressão linear generalizada foram escolhidos devido à simplicidade de sua estrutura, a qual permite uma interpretação direta e fácil da relação entre as propriedades do solo e os atributos de terreno. Esse aspecto é de suma importância, uma vez que as FPESe não devem ser vistas apenas como formas de predizer propriedades do solo, mas também de entender os processos e características ambientais que governam essas propriedades. Além disso, o aumento da complexidade das FPESe devido à introdução de variáveis explanatórias adicionais e a ajustes de ordem mais elevada os tornam menos generalizáveis e, portanto, menos úteis (Gessler et al., 1996). É por isso que Minasny e McBratney (2007) afirmam que a melhoria nas predições de propriedades do solo não depende tanto do aumento da complexidade dos modelos matemáticos e dos métodos estatísticos utilizados, mas sim na utilização de dados mais úteis e de maior qualidade.

A segunda possibilidade em termos de modelos matemáticos é o uso daqueles conhecidos como híbridos. Esses modelos combinam métodos geoestatísticos e modelos de regressão linear *(regression-kriging).* Nesse caso, o modelo de regressão linear seria utilizado

para estimar o componente determinístico da distribuição do tamanho de partículas do solo, enquanto a krigagem seria utilizada para estimar o componente aleatório. (Mendonça-Santos et al., 2010) obtiveram resultados satisfatórios através desse método ao mapear os estoques de carbono nos solos do estado do Rio de Janeiro. O inconveniente desse procedimento é que a área a ser mapeada necessita, obrigatoriamente, ser amostrada para que seja possível interpolar os resíduos das predições realizadas através do modelo de regressão linear.

8 CONCLUSÕES

Funções de predição espacial de solos (FPESe) podem ser construídas com atributos de terreno para estimar a distribuição do tamanho de partículas do solo em áreas de geologia complexa como a região do Rebordo do Planalto do RS. Entretanto, a heterogeneidade geológica reduz a capacidade preditiva das FPESe construídas para algumas das frações de tamanho de partícula. Sobretudo se o material de origem for o fator de formação do solo com maior efeito sobre a distribuição do tamanho de partículas do solo como ocorre na região do Rebordo do Planalto do RS.

Em áreas de maior homogeneidade geológica os atributos de terreno são mais bem correlacionados com a distribuição do tamanho de partículas do solo. Entretanto, a estratificação da paisagem de acordo com o tipo do material de origem (ígnea extrusiva ou sedimentar) não é uma garantia de que a capacidade preditiva das FPESe será aumentada. Sobretudo se a diversidade de rochas sedimentares e ígneas extrusivas for elevada, como ocorre na região do Rebordo do Planalto do RS. Como não existem mapas geológicos acurados, a geração de tal informação é fundamental para a melhoria da capacidade preditiva das FPESe.

Como o material de origem é o fator de formação dos solos com maior influência sobre a distribuição do tamanho de partículas, o relevo acaba exercendo influência secundária. Sua influência se dá pelo controle da erosão, sobretudo em canais, com maior peso para os processos de perda e transferência das partes mais elevadas, seguido do processo de acumulação nas partes mais baixas da paisagem.

Os atributos de terreno classificados como primários estão entre os mais importantes para a construção de FPESe porque explicam a maior parte da variância predita da distribuição do tamanho de partículas do solo. Mas isso se deve ao fato de que a elevação (ELEV) está intimamente relacionada ao material de origem na região do Rebordo do Planalto do RS. A importância dos atributos secundários nas FPESe construídas é pequena porque a influência do relevo sobre a distribuição do tamanho de partículas também é pequena. Mas seu uso em substituição a múltiplos atributos primários elimina quase que completamente a multicolinearidade entre as variáveis preditoras.

9 REFERÊNCIAS

AITCHISON, J. **A concise guide to compositional data analysis.** CDA workshop, Girona, Espanha. 2003. 134p.

AITCHISON, J. The statistical analysis of compositional data. **Journal of the Royal Statistical Society. Series B (Methodological)**, v.44, p.139-177, 1982.

BASHER, L.R. Is pedology dead and buried? **Australian Journal of Soil Research**, v.35, p.979-994, 1997.

BERNOUX, M.; ARROUAYS, D.; CERRI, C.; CERRI, C. Regional organic carbon storage maps of the western brazilian amazon based on prior soil maps and geostatistical interpolation. In: LAGACHERIE, P.; MCBRATNEY, A.B.; VOLTZ, M. (Eds.). **Digital soil mapping - an introductory perspective.** Amsterdam: Elsevier, 2006. p.497-506.

BEVEN, K.; KIRKBY, N. A physically based variable contributing area model of basin hydrology. **Hydrological Sciences Bulletin**, v.24, p.43-69, 1978.

BISHOP, T.; MCBRATNEY, A.B. A comparison of prediction methods for the creation of field-extent soil property maps. **Geoderma**, v.103, p.149-160, 2001.

BISHOP, T.; MINASNY, B. Digital soil-terrain modeling: the predictive potential and uncertainty. In: GRUNWALD, S. (Ed.). **Environmental soil-landscape modeling - geographic information technologies and pedometrics**. Boca Raton: Taylor and Francis, 2006. p.185-213.

BÖHNER, J.; KÖTHE, R.; CONRAD, O.; GROSS, J.; RINGELER, A.; SELIGE, T. Soil regionalisation by means of terrain analysis and process parameterisation. In: MICHELI, E.; NACHTERGAELE, F.O.; JONES, R.J.; MONTANARELLA, L. (Eds.). **Soil Classification 2001**. European Soil Bureau, Research Report No. 7, EUR 20398 EN, Office for Official Publications of the European Communities, Luxembourg, 2002. p.213-222.

BÖHNER, J.; SELIGE, T. Spatial prediction of soil attributes using terrain analysis and climate regionalisation. In: BÖHNER, J.; MCCLOY, K.; STROBL, J. (Eds.). **SAGA – analysis and modelling applications**. Göttingen: Göttinger Geographische Abhandlungen, 2006. p.13-28.

BRASIL. **Mapa geológico da folha Santa Maria – RS**. Santa Maria: Universidade Federal de Santa Maria, Centro de Ciências Naturais e Exatas, Departamento de Geociências, 1980. 1 mapa, p&b. Escala 1:50.000.

BROWN, D.J. Using a global VNIR soil-spectral library for local soil characterization and landscape modeling in a 2nd-order Uganda watershed. **Geoderma**, v.140, p.444-453, 2007.

BRUM, A. J. **Modernização da agricultura: trigo e soja**. Ijuí: Vozes, 1988. 200p.

BRUS, D.; DE GRUIJTER, J. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). **Geoderma**, v.80, p.1-44, 1997.

BRUS, D.J.; HEUVELINK, G.B. Optimization of sample patterns for universal kriging of environmental variables. **Geoderma**, v.138, p.86-95, 2007.

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

BUCCIANTI, A.; MATEU-FIGUERAS, G.; PAWLOWSKY-GLAHN, V. **Compositional data analysis in the geosciences: from theory to practice**. London: Geological Society of London. 2006. 213p.

CAMARGO, E.C. **Desenvolvimento, implementação e teste de procedimentos geoestatísticos (krigeagem) no sistema de processamento de informações georreferenciadas (SPRING)**. 1997. 146f. Dissertação (Mestrado em Sensoriamento Remoto) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, SP. 1997.

CAMARGO, F.A.; INDA JUNIOR, A. Programa de Pós-graduação em Ciência do Solo - Universidade Federal do Rio Grande do Sul. **Boletim Informativo da SBCS**, v.36, p.57-60, 2011.

CAMBARDELLA, C.A.; MOORMAN, T.B.; NOVAK, J.M.; PARKIN, T.B.; KARLEN, D.L.; TURCO, R.F.; KONOPKA, A.E. Field-scale variability of soil properties in central Iowa soils. **Soil Science Society of America Journal**, v.58, p.1501-1511, 1994.

CARRÉ, F.; MCBRATNEY, A.B.; MAYR, T.; MONTANARELLA, L. Digital soil assessments: beyond DSM. **Geoderma**, v.142, p.69-79, 2007.

CONRAD, O. **Ableitung hydrologisch relevanter Reliefparameter aus einem Digitalen Geländemodell (am Beispiel des Einzugsgebietes Linnengrund / Kaufunger Wald)**. 1998. 95f. Dissertação (Mestrado em Geografia) – Universidade de Göttingen, Göttingen, Alemanha, 1998.

COORDENAÇÃO DE APERFEIÇOAMENTO DE PESSOAL DE NÍVEL SUPERIOR – CAPES. **Tabela de áreas do conhecimento**. 2011. Disponível em: <http://www.capes.gov.br /avaliacao/tabela-de-areas-de-conhecimento>. Acesso em: 26 dez. 2011 .

DALMOLIN, R.S.D.; PEDRON, F.A. Solos do município de Santa Maria. **Ciência e Ambiente**, v.38, p.59-77, 2009.

DELIGNETTE-MULLER, M. L.; POUILLOT, R.; DENIS, J.-B.; DUTANG, C. ***fitdistrplus*: help to fit of a parametric distribution to non-censored or censored data**. 2010. R

package version 0.1-3. Disponível em: < http://cran.r-project.org/web/packages/fitdistrplus /index.html >.

DIAS, C. **Mapeamento digital de solos se organiza no Brasil**. 2011. Disponível em: <http://www.cnps.embrapa.br/noticias/banco_noticias/20111130.html>. Acesso em: 15 dez. 2011.

DIAS, J.R. **Aplicação do modelo hidrológico AGNPS2001 utilizando dados observados na bacia do arroio Vacacaí-Mirim.** 2003. Dissertação (Mestrado em Engenharia Civil) - Universidade Federal de Santa Maria, Santa Maria, RS. 2003.

DIETZ, T.; ROSA, E.A.; YORK, R. Driving the human ecological footprint. **Frontiers in Ecology and the Environment**, v.5, p.13-18, 2007.

DILL, P.; PAIVA, E.; PAIVA, J.; ROCHA, J. Assoreamento do reservatório do Vacacaí-Mirim e sua relação com a deterioração da bacia hidrográfica contribuinte. **Revista Brasileira de Recursos Hídricos**, v.9, p.7-15, 2004.

EDUCATION FIRST – EF. **Índice de proficiência em inglês da EF**. Cambridge: Education First – EF, 2011. 21p.

EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA – EMBRAPA. **Handbook of methods of soil analysis**, 2.ed. Rio de Janeiro: Embrapa, 1997.

ENVIRONMENTAL SYSTEMS RESEARCH INSTITUTE – ESRI. **ArcMap 9.3.** Redlands, 2009.

ENVIRONMENTAL SYSTEMS RESEARCH INSTITUTE – ESRI. **Using Topo to Raster in 3D Analyst.** 2009. Disponível em: <http://webhelp.esri.com/arcgiSDEsktop/9.3/index. cfm?TopicName=Using%20topo%20to%20raster%20in%203d%20analyst>; Acesso em: 30 Ago. 2011.

ESPINDOLA, C.R. **Retrospectiva crítica sobre a pedologia - um repasse biliográfico**. Campinas: Editora da Unicamp, 2008. 397p.

FARIA, C. **FMI e a dívida externa brasileira**. 2007. Disponível em: <http://www.infoescola.com/geografia/fmi-e-a-divida-externa-brasileira/>. Acesso em: 27 dez. 2011.

FEYERABEND, P. **Contra o método**. Rio de Janeiro: Livraria Franscisco Alves Editora S.A. 1977. 188p.

FINKE, P.A. On digital soil assessment with models and the Pedometrics agenda. **Geoderma**, v.171-172, p.3-15, 2012.

FOOD AND AGRICULTURE ORGANIZATION – FAO. **Guidelines for soil description**. Roma: Food and Agriculture Organization of The United Nations, 2006.

FOOD AND AGRICULTURE ORGANIZATION – FAO. **The FAO voluntary guidelines for the right to food: lasting solutions against hunger**. 2005. Disponível em: <http://www.fao.org/righttofood/KC/downloads/vl/docs/ AH269_en.pdf>. Acesso em: 27 dez. 2011.

FREEMAN, G. Calculating catchment area with divergent flow based on a regular grid. **Computers and Geosciences**, v.17, p.413-422, 1991.

GATIBONI, L.C. Pós-graduação em Manejo do Solo - Universidade do Estado de Santa Catarina, campus Lages. **Boletim Informativo da SBCS**, v.36, p.53-56, 2011.

GESSLER, P.; MCKENZIE, N.; HUTCHINSON, M. Progress in soil-landscape modeling and spatial prediction of soil attributes for environmental models. In: INTERNATIONAL CONFERENCE ON INTEGRATING GIS AND ENVIRONMENTAL MODELING, 6., 1996, Sante Fe, New Mexico. **Anais…** Santa Barbara: National Center for Geographic Information and Analysis, 1996. Discponível em: < http://www.ncgia.ucsb.edu/conf/ SANTA_FE_CD-ROM/sf_papers/gessler_paul/my_paper.html >. Acesso em 20 Ago. 2011.

GESSLER, P.E.; MOORE, I.D.; MCKENZIE, N.J.; RYAN, P.J. Soil-landscape modelling and spatial prediction of soil attributes. **International Journal of Geographical Information Systems**, v.9, p.421-432, 1995.

GIASSON, E.; CLARKE, R.T.; INDA JUNIOR, A.V.; MERTEN, G.H.; TORNQUIST, C.G. Digital soil mapping using multiple logistic regression on terrain parameters in southern Brazil. **Scientia Agricola**, v.63, p.262-268, 2006.

GLEISER, M. **Criação imperfeita**. Rio de Janeiro: Editora Record, 2010. 368p.

GLOBALSOILMAP.NET. **Specifications Version 1 GlobalSoilMap.net products**, 2011. Disponível em: < http://www.globalsoilmap.net/system/files/GlobalSoilMap_net_ specifications_v2_0_edited_draft_Sept_2011_RAM_V12.pdf >. Acesso em: 22 dez. 2011.

GOBIN, A.; CAMPLING, P.; FEYEN, J. Soil-landscape modelling to quantify spatial variability of soil texture. **Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere**, v.26, p.41-45, 2001.

GRUBER, S.; PECKHAM, S. Land-surface parameters and objects in hydrology. In: HENGL, T.; REUTER, H.I. (Eds.). **Geomorphometry - Concepts, Software, Applications.** Amsterdam: Elsevier, 2009. p.171-194.

GRUWALD, S. Multi-criteria characterization of recent digital soil mapping and modeling approaches. **Geoderma**, v.152, p.195-207, 2009.

GRUNWALD, S. What do we really know about the space–time continuum of soil-landscapes? In: GRUNWALD, S. (Ed.). **Environmental soil-landscape modeling - geographic information technologies and pedometrics.** Boca Raton: Taylor and Francis, 2006. p.3-36.

HAIR, J.F.; BLACK, B.; BABIN, B.; ANDERSON, R.E. **Multivariate data analysis**. New Jersey: Pearson Prentice Hall, 2010. 760p.

HARTEMINK, A. E.; MCBRATNEY, A.B. A soil science renaissance. **Geoderma**, v.148, p.123-129, 2008.

HEIBERGER, R.M. **HH: Statistical Analysis and Data Display: Heiberger and Holland**. 2011.

HEIM, A.; WEHRLI, L.; EUGSTER, W.; SCHMIDT, M. Effects of sampling design on the probability to detect soil carbon stock changes at the Swiss CarboEurope site Lägeren. **Geoderma**, v.149, p.347-354, 2009.

HENGL, T.; MACMILLAN, R. Geomorphometry - a key to landscape mapping and modelling. In: HENGL, T.; REUTER, H.I. (Eds.). **Geomorphometry - concepts, software, applications**. Amsterdam: Elsevier, 2009. p.433-460.

HENGL, T.; ROSSITER, D.G. Supervised landform classification to enhance and replace photo-interpretation in semi-detailed soil survey. **Soil Science Society of America Journal**, v.67, p.1810-1822, 2003.

HEUVELINK, G. The definition of pedometrics. **Pedometron**, v.15, p.11-12, 2003.

HILLEL, D. **Encyclopedia of soils in the environment**. Oxford: Elsevier, 2005. 2200p.

HUIJBREGTS, C. Regionalized variables and quantitative analysis of spatial data. In: DAVIS, J.; MCCULLAGH, M. (Eds.). **Display and analysis of spatial data**. New York: John Wiley, 1975. p.38-53.

HUTCHINSON, M.F. A locally adaptive approach to the interpolation of digital elevation models. In: INTERNATIONAL CONFERENCE ON INTEGRATING GIS AND ENVIRONMENTAL MODELING, 6., 1996, Sante Fe, New Mexico. **Anais…** Santa Barbara: National Center for Geographic Information and Analysis, 1996. Discponível em: < http://www1.gsi.go.jp/geowww/globalmap-gsi/gtopo30/papers/local.html>. Acesso em 20 Ago. 2011.

HUTCHINSON, M.F. A new procedure for gridding elevation and stream line data with automatic removal of spurious pits. **Journal of Hydrology**, v.106, p.211-232, 1989.

HUTCHINSON, M.F.; GALLANT, J.C. Digital elevation models and representation of terrain shape. In: WILSON, D.J.; GALLANT, J.C. (Eds.). **Terrain analyis: principles and applications**. New York: John Wiley and Sons, 2000. p.29-50.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE. **Censo demográfico 2010**. 2010. Disponível em: <http://www.censo2010.ibge.gov.br/>. Acesso em: 27 dez. 2011.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE. **Vocabulário básico de recursos naturais e meio ambiente**, 2.ed. Rio de Janeiro: IBGE. 2004. 332p.

JENNY, H. **Factors of soil formation - a system of quantitative pedology**. New York: Dover Publications, 1941. 281p.

KER, J. C. O futuro da pedologia no Brasil. **Boletim Informativo da SBCS**, v.25, p.18-21. 1999.

KER, J. C.; NOVAIS, R. F. Fundamentos para o desenvolvimento da pedologia e da fertilidade do solo. In: CONGRESSO BRASILEIRO DE CIÊNCIA DO SOLO, 24., 2003, Ribeirão Preto. **Anais...** Viçosa: Sociedade Brasileira de Ciência do Solo, 2003. Disponível em: <http://jararaca.ufsm.br/websites/classolos/download/TextosSol/Texto03.pdf>. Acesso em: 27 dez. 2011.

KERRY, R.; OLIVER, M. Comparing sampling needs for variograms of soil properties computed by the method of moments and residual maximum likelihood. **Geoderma**, v.140, p.383-396, 2007.

KITAMURA, P. Agricultura e desenvolvimento sustentável: uma agenda para discussão. **Ciência e Ambiente**, v.4, p.37-49, 1993.

KÖETHE, R.; LEHMEIER, F. **SARA, System zur Automatischen Relief-Analyse - Benutzerhandbuch**, 2.ed. Universidade de Göttingen, Instituto de Geografia, Göttingen, Alemanha. 1996. (não publicado).

KRASILNIKOV, P.; ARNOLD, R. Soil classifications and their correlations. In: KRASILNIKOV, P.; MARTÍ, J.-J. I.; ARNOLD, R.; SHOBA, S. (Eds.). **A handbook of soil terminology, correlation and classification**. London: Earthscan, 2009. p.45-336.

KRASILNIKOV, P.; IBÁÑEZ, J.-J.; ARNOLD, R. The theoretical bases of soil classifications. In: KRASILNIKOV, P.; MARTÍ, J.-J. I.; ARNOLD, R.; SHOBA, S. (Eds.). **A handbook of soil terminology, correlation and classification.** London: Earthscan, 2009. p.5-43.

KUHN, M.; WING, J.; WESTON, S.; WILLIAMS, A.; KEEFER, C.; ENGELHARDT, A. **caret: Classification and Regression Training**. 2012.

KUTNER, M.H.; NACHTSHEIM, C.J.; NETER, J.; LI, W. **Applied linear statistical models.** 5.ed. New York: McGraw-Hill, 2004.

LAGACHERIE, P.; LEGROS, J.; BURFOUGH, P. A soil survey procedure using the knowledge of soil pattern established on a previously mapped reference area. **Geoderma**, v.65, p.283-301, 1995.

LAGACHERIE, P.; MCBRATNEY, A.B. Spatial soil information systems and spatial soil inference systems: perspectives for digital soil mapping. LAGACHERIE, P.; MCBRATNEY, A.B.; VOLTZ, M. (Eds.). **Digital soil mapping - an introductory perspective.** Amsterdam: Elsevier, 2006. p.3-22.

LAGACHERIE, P.; ROBBEZ-MASSON, J.; NGUYEN-THE, N.; BARTHÈS, J. Mapping of reference area representativity using a mathematical soilscape distance. **Geoderma**, v.101, p.105-118, 2001.

LAL, R.; KIMBLE, J.M.; FOLLETT, R.F. Pedospheric processes and the carbon cycle. In: LAL, R.; KIMBLE, J.M.; FOLLETT, R.F.; STEWART, B.A. (Eds.). **Soil processes and the carbon cycle**. Boca Raton: CRC Press, 1997. p.1-8.

LARK, R.; BISHOP, T. Cokriging particle size fractions of the soil. **European Journal of Soil Science**, v.58, p.763-774, 2007.

LARK, R.; BISHOP, T.; WEBSTER, R. Using expert knowledge with control of false discovery rate to select regressors for prediction of soil properties. **Geoderma**, v.138, p.65-78, 2007.

LEEPER, G.W. The classification of soils. **Journal of Soil Science**, v.7, p.59-64, 1956.

MACARINI, J.P. A política econômica do governo Médici: 1970-1973. **Nova Economia**, v.15, p.53-92, 2005.

MACIEL FIHO, C.L. **Carta geotécnica de Santa Maria**. Santa Maria: Imprensa Universitária – UFSM, 1990. 21p.

MACIEL FILHO, C.; SARTORI, P.; VEIGA, P.; GASPARETTO, N. **Mapa Geológico da Folha de Camobi. Texto explicativo**. Santa Maria: Universidade Federal de Santa Maria, 1988. 10p.

MALUF, J. A new climatic classification for the state of Rio Grande do Sul, Brazil. **Revista Brasileira de Agrometeorologia**, v.8, p.141-150, 2000.

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

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

MCBRATNEY, A.B.; MINASNY, B.; TRANTER, G. Necessary meta-data for pedotransfer functions. **Geoderma**, v.160, p.627-629, 2011.

MCBRATNEY, A.B.; ODEH, I.O.; BISHOP, T.F.; DUNBAR, M.S.; SHATAR, T.M. An overview of pedometric techniques for use in soil survey. **Geoderma**, v.97, p.293-327, 2000.

MCDONALD, B. A teaching note on Cook's distance – a guideline. **Research Letters in the Information and Mathematical Sciences**, v.3, p.127-128, 2002.

MCKENZIE, N.; GESSLER, P. E.; RYAN, P. J.; O'CONNELL, D. The role of terrain analysis in soil mapping. In: WILSON, J.; GALLANT, J. (Eds.). **Terrain analysis: principles and applications**, New York: John Wiley and Sons, 2000. p.245-266.

MCKENZIE, N.; GRUNDY, M.; WEBSTER, R.; RINGROSE-VOASE, A. **Guidelines for surveying soil and land resources**. 2.ed. Melbourne: CSIRO Publishing, 2008. 576p.

MELLO, C. R.; ÁVILA, L. F.; NORTON, L. D.; SILVA, A. M.; MELLO, J. M.; BESKOW, S. Spatial distribution of top soil water content in an experimental catchment of Southeast Brazil. **Scientia Agricola**, v.68, p.285-294, 2011.

MENDONÇA-SANTOS, M. L.; DART, R.; SANTOS, H.; COELHO, M.; BERBARA, R.; LUMBRERAS, J. Digital soil mapping of topsoil organic carbon content of Rio de Janeiro State, Brazil. In: BOETTINGER, J. L.; HOWELL, D. W.; MOORE, A. C.; HARTEMINK, A. E.; KIENAST-BROWN, S. (Eds.). **Digital soil mapping: bridging research, environmental application, and operation**, New York: Springer, 2010. p.255-265.

MENDONÇA-SANTOS, M. L.; SANTOS, H. G. **Mapeamento digital de classes e atributos de solos - métodos, paradigmas e novas técnicas**. Rio de Janeiro: Embrapa Solos, 2003. 17p.

MENDONÇA-SANTOS, M. L.; SANTOS, H. The state of the art of Brazilian soil mapping and prospects for digital soil mapping. In: LAGACHERIE, P.; MCBRATNEY, A. B.; VOLTZ, M. (Eds.). **Digital soil mapping - an introductory perspective**. Amsterdam: Elsevier, 2006. p.39-54.

MIGUEL, P. **Pedological characterization, land use and modeling of the soil loss in hillslope areas the Plateau Border of RS**. 2010. 112p. Dissertação (Mestrado em Ciência do Solo) - Universidade Federal de Santa Maria, Santa Maria, RS. 2010.

MILNE, A.; LARK, M. Finding the boundary. **Pedometron**, v.25, p.10-13, 2008.

MINASNY, B.; MCBRATNEY, A.B. Spatial prediction of soil properties using EBLUP with the Matérn covariance function. **Geoderma**, v.140, p.324-336, 2007.

MINELLA, J.P.G.; MERTEN, G.H.; RUHOFF, A.L. Use of spatial representation to calculate the topographic factor in the revised universal soil loss equation in watersheds. **Revista Brasileira de Ciência do Solo**, v.34, p.1455-1462, 2010.

MONTEIRO, C.A. A dimensão da pobreza, da desnutrição e da fome no Brasil. **Estudos Avançados**, v.17, p.7-20, 2003.

MOORE, I.D.; BURCH, G.; MACKENZIE, D. Topographic effects on the distribution of surface soil water and the location of ephemeral gullies. **Transactions of the ASAE**, v.31, p.1098-1107, 1988.

MOORE, I.D.; GESSLER, P.E.; NIELSEN, G.A.; PETERSON, G.A. Soil attribute prediction using terrain analysis. **Soil Science Society of America Journal**, v.57, p.443-452, 1993.

MOORE, I.D.; GRAYSON, R.; LADSON, A. Digital terrain modelling: a review of hydrogical, geomorphological, and biological applications. **Hydrological Processes**, v.5, p.3-30, 1991.

MORIN, E. **Para sair do século XX - as grandes questões do nosso tempo**. Rio de Janeiro: Nova Fronteira, 1986. 361 p.

MOYES, J.; SHANGGUAN, W. ***soiltexture*: functions for soil texture plot, classification and transformation**. 2011. R package version 1.2.2. Disponível em: <http://CRAN.R-project.org/package=soiltexture>.

NEUMANN, P. **The impact of the fragmentation and of the lands' format in the family farm systems**. 2003. 326p. Tese (Doutorado em Engenharia de Produção) - Universidade Federal de Santa Catarina, Florianópolis, SC. 2003.

ODEH, I. O.; TODD, A. J.; TRIANTAFILIS, J. Spatial prediction of soil particle-size fractions as compositional data. **Soil Science**, v.168, p.501-515, 2003.

OLAYA, V. **A gentle introduction to SAGA GIS**. 2004. 202p. Disponível em: <http://ufpr.dl.sourceforge.net/project/saga-gis/SAGA%20-%20Documentation/SAGA%20Documents/SagaManual.pdf>.

OLAYA, V. Basic land-surface parameters. In: HENGL, T.; REUTER, H.I. (Eds.). **Geomorphometry - Concepts, Software, Applications.** Amsterdam: Elsevier, 2009. p.141-169.

OLAYA, V.; CONRAD, O. Geomorphometry in SAGA. In: HENGL, T.; REUTER, H.I. (Eds.). **Geomorphometry - Concepts, Software, Applications**. Amsterdam: Elsevier, 2009. p.293-308.

OLIVEIRA JUNIOR, J. C. Spatial variability of mineralogical properties in soil of the Guabirotuba formation of Curitiba (PR). **Revista Brasileira de Ciência do Solo**, v.35, p.1481-1490, 2011.

PARK, S.; VLEK, P. Environmental correlation of three-dimensional soil spatial variability: a comparison of three adaptive techniques. **Geoderma**, v.109, p.117-140, 2002.

PAWLOWSKY-GLAHN, V.; EGOZCUE, J. J. Compositional data and their analysis: an introduction. In: BUCCIANTI, A.; MATEU-FIGUERAS, G.; PAWLOWSKY-GLAHN, V. (Eds.). **Compositional data analysis in the geosciences: from theory to practice**. London: Geological Society, 2006. p.1-10.

PEDRON, F.A. **Mineralogia, morfologia e classificação de saprolitos e Neossolos derivados de rochas vulcânicas no Rio Grande do Sul**. 2007. 160f. Tese (Doutorado em Ciência do Solo) – Universidade Federal de Santa Maria, Santa Maria, RS. 2007.

PINHEIRO, R.J.; SOARES, J.M. Condicionantes geológicos-geotécnicos de movimentos de massa na encosta da Serra Geral - RS. **Teoria e Prática na Engenharia Civil**, v.4, p.59-68, 2004.

QUINN, P.; BEVEN, K.; CHEVALLIER, P.; PLANCHON, O. The prediction of hillslope flow paths for distributed hydrological modelling using digital terrain models. **Hydrological Processes**, v.5, p.59-79, 1991.

R DEVELOPMENT CORE TEAM. R: **A language and environment for statistical computing**. 2011. R Foundation for Statistical Computing, Vienna, Áustria. Disponível em: <http://www.R-project.org/>.

RAMOS, D.P. **Desafios da pedologia brasileira frente ao novo milênio**. 2003. Palestra realizada no XXIX Congresso Brasileiro de Ciência do Solo. Ribeirão Preto, SP, Julho 2003. Disponível em: < http://jararaca.ufsm.br/websites/dalmolin/download/textospl/desafio.pdf >. Acesso em: 27. dez. 2011.

RAWLINS, B.G.; WEBSTER, R.; TYE, A.M.; LAWLEY, R.; O'HARA, S.L. Estimating particle-size fractions of soil dominated by silicate minerals from geochemistry. **European Journal of Soil Science**, v.60, p.116-126, 2009.

REFAEILZADEH, P.; TANG, L.; LIU, H. Cross validation. In: ÖZSU, M.T.; LIU, L. (Eds.). **Encyclopedia of database systems**. New York: Springer, 2009.

REICHERT, J.M.; REINERT, D.J.; BRAIDA, J.A. Qualidade dos solos e sustentabilidade de sistemas agrícolas. **Ciência & Ambiente**, v.27, p.29-48, 2003.

REVELLE, W. ***psych*: procedures for psychological, psychometric, and personality research**. 2011. R package version 1.0-97. Disponível em: < http://cran.r-project.org/web/packages/psych/index.html >.

RILEY, S.J.; DEGLORIA, S.D.; ELLIOT, R. A terrain ruggedness index that quantifies topographic heterogeneity. **Intermountain Journal of Sciences**, v.5, p.1-4, 1999.

ROECKER, S.; THOMPSON, J. Scale effects on terrain attribute calculation and their use as environmental covariates for digital soil mapping. In: BOETTINGER, J.L.; HOWELL, D. W.; MOORE, A. C.; HARTEMINK, A. E.; KIENAST-BROWN, S. (Eds.). **Digital soil**

mapping: bridging research, environmental application, and operation. New York: Springer, 2010. p.55-66.

RUSSELL, B. How to become a man of genius. **Hearst newspapers**.1932. Disponível em: <http://www.intelectu.com/intelectu_archive_win_07_09.html>. Acesso em: 20 dez. 2011.

SAGA DEVELOPMENT TEAM. **SAGA GIS**. Hamburg, Alemanha, 2010. Versão 2.0.6. Disponível em: <http://www.saga-gis.org, 2010>.

SAMUEL-ROSA, A.; DALMOLIN, R.S.D.; PEDRON, F.A. Caracterização do solo de cobertura de aterros encerrados com ferramentas (geo)estatísticas. **Engenharia Sanitaria e Ambiental**, v.16, p.121-126.

SAMUEL-ROSA, A.; MIGUEL, P.; DALMOLIN, R.S.D.; PEDRON, F.A. Uso da terra no Rebordo do Planalto do Rio Grande do Sul. **Ciência e Natura**, v.33, p.161-173, 2011.

SANCHEZ, P. A.; AHAMED, S.; CARRÉ, F.; HARTEMINK, A. E.; HEMPEL, J.; HUISING, J.; LAGACHERIE, P.; MCBRATNEY, A. B.; MCKENZIE, N. J.; MENDONÇA-SANTOS, M. L.; MINASNY, B.; MONTANARELLA, L.; OKOTH, P.; PALM, C. A.; SACHS, J. D.; SHEPHERD, K. D.; VÅGEN, T.-G.; VANLAUWE, B.; WALSH, M. G.; WINOWIECKI, L. A.; ZHANG, G.-L. Digital soil map of the world. **Science**, v.325, p.680-681, 2009.

SANTOS, H.G.; JACOMINE, P.K.T.; ANJOS, L.H.C.; OLIVEIRA, V.A.; OLIVEIRA, J.B.; COELHO, M.R.; LUMBRERAS, J.F.; CUNHA, T.J.F. **Sistema Brasileiro de Classificação de Solos.** 2.ed. Rio de Janeiro: EMBRAPA, 2006. 306p.

SARTORI, P. Geologia e geomorfologia de Santa Maria. **Ciência e Ambiente**, v.38, p.19-42, 2009.

SILVA, L.K. A migração dos trabalhadores gaúchos para a Amazônia Legal (1970-1985). II - A política de ocupação das fronteiras amazônicas. **Klepsidra - Revista Virtual de História**, 2005. Disponível em: <http://www.klepsidra.net/klepsidra24/agro-rs2.htm>. Acesso em: 27 dez. 2011.

SILVEIRA, F.G.; ALMEIDA, M.E. Fome, produção alimentar e distribuição da renda. **Indicadores Econômicos FEE**, v.19, p.151-166, 1992.

SKIDMORE, T. **Brasil: de Castelo a Tancredo**. Rio de Janeiro: Ed. Paz e Terra, 1989. 608p.

STRECK, E.V.; KÄMPF, N.; DALMOLIN, R.S.D.; KLAMT, E.; NASCIMENTO, P.C.; SCHNEIDER, P.; GIASSON, E.; PINTO, L.F.S. **Solos do Rio Grande do Sul**. 2.ed. Porto Alegre: Emater/RS, 2008. 222p.

SUMFLETH, K.; E DUTTMANN, R. Prediction of soil property distribution in paddy soil landscapes using terrain data and satellite information as indicators. **Ecological Indicators**, v.8, p.485-501, 2008.

TEN CATEN, A. M**apeamento digital de solos: metodologias para atender a demanda por informação espacial em solos**. 2011. 108f. Tese (Doutorado em Ciência do Solo) - Universidade Federal de Santa Maria, Santa Maria, RS. 2011.

TEN CATEN, A.; DALMOLIN, R.S.D.; PEDRON, F.A.; MENDONÇA-SANTOS, M.L. Componentes principais como preditores no mapeamento digital de classes de solos. **Ciência Rural**, v.41, p.1170-1176, 2011a.

TEN CATEN, A.; DALMOLIN, R.S.D.; PEDRON, F.A.; MENDONÇA-SANTOS, M.L. Estatística multivariada aplicada à diminuição do número de preditores no mapeamento digital do solo. **Pesquisa Agropecuária Brasileira**, v.46, p.554-562, 2011b.

TEN CATEN, A.; DALMOLIN, R.S.D.; PEDRON, F.A.; MENDONÇA-SANTOS, M.L. Extrapolação das relações solo-paisagem a partir de uma área de referência. **Ciência Rural**, v.41, p.812-816, 2011c.

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

TROEH, F. R. Landform parameters correlated to soil drainage. **Soil Science Society of America Journal**, v.28, p.808-812, 1964.

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO – UFRRJ. **Edital nº 15, de 4 de maio de 2011.** Torna público que estarão abertas as inscrições para Concurso Público de Provas e Títulos para ingresso na Carreira do Magistério Superior na Área de Concentração de Mapeamento Digital de Solos e Pedologia Aplicada. Diário Oficial da União, 5 mai. 2011. Disponível em: <http://www.in.gov.br/imprensa/visualiza/index.jsp?jornal=3epagina=60e data=05/05/2011>. Acesso em: 15 dez. 2011.

VARMUZA, K.; FILZMOSER, P. **Introduction to multivariate statistical analyis in chemometrics**. Boca Raton, Florida, USA: CRC Press (Taylor & Francis), 2009.

VENABLES, W.N.; RIPLEY, B.D. **Modern applied statistics with S.** 4.ed. New York: Springer, 2002. 495p.

VIER, S. **Analfabetismo funcional chega aos bancos escolares**. 2010. Disponível em: <http://www.brasilwiki.com.br/noticia.php?id_noticia=31428>. Acesso em 27 dez. 2011.

VOLTZ, M.; LAGACHERIE, P.; LOUCHART, X. Predicting soil properties over a region using sample information from a mapped reference area. **European Journal of Soil Science**, v.48, p.19-30, 1997.

WEBSTER, R. Statistics to support soil research and their presentation. **European Journal of Soil Science**, v.52, p.331-340, 2001.

WEBSTER, R.; BURROUGH, P. A. Computer-based soil mapping of small areas from sample data. **Journal of Soil Science**, v.23, p.210-221, 1972.

WEBSTER, R.; OLIVER, M. **Statistical methods in soil and land resource survey**. Oxford: Oxford University Press, 1990. 316p.

WEISBERG, S. **Applied linear regression**. 3.ed. New Jersey: John Wiley and Sons, 2005.

WIKIPEDIA. **Ciências do solo**. 2011. Disponível em: <http://pt.wikipedia.org/wiki/Ci%C3%AAncias_do_solo>. Acesso em: 28 dez. 2011.

WIKIPEDIA. **Coeficiente de determinação**. 2012a. Disponível em: <http://pt.wikipedia.org/wiki/R%C2%B2>. Acesso em: 13 Fev. 2012.

WIKIPEDIA. **Mean squared error**. 2012b. Disponível em: <http://en.wikipedia.org/wiki/Mean_squared_error>. Acesso em: 13 Fev. 2012.

WIKIPEDIA. **Root-mean-square deviation**. 2012c. Disponível em: <http://en.wikipedia.org/wiki/Root_mean_square_deviation>. Acesso em: 13 Fev. 2012.

WILSON, J.P.; GALLANT, J.C. Digital terrain analysis. In: WILSON, J.P.; GALLANT, J.C. (Eds.). **Terrain analyis: principles and applications**. New York: John Wiley and Sons, 2000a. p.1-27.

WILSON, J.; GALLANT, J. **Terrain analysis: principles and applications**, New York: John Wiley and Sons, 2000b. 485p.

WOOD, E.F.; SIVAPALAN, M.; BEVEN, K. Similarity and scale in catchment storm response. **Reviews of Geophysics**, v.28, p.1-18, 1990.

ZATTI, V. Nietzsche e a educação. In: OLIVEIRA, W.R.; SILVA, S.S (Org.). **Leituras em Educação**. São Mateus: Opção, 2008, v.2, p.101-116.

ZEVENBERGEN, L.; THORNE, C. Quantitative analysis of land surface topography. **Earth Surface Processes and Landforms**, v.12, p.47-56, 1987.

10 ANEXOS

Anexo 1 – Questionário sobre mapeamento digital de solos no Brasil

a) Enviado aos autores de trabalhos publicados

Questionário sobre publicações abordando o mapeamento digital de solos no Brasil

Caro pedometrista

Meu nome é Alessandro Samuel-Rosa, aluno do Programa de Pós-Graduação em Ciência do Solo da Universidade Federal de Santa Maria, sob a orientação do Prof. Dr. Ricardo Simão Diniz Dalmolin. O tema de meu projeto de pesquisa é mapeamento digital de solo (MDS). Durante a revisão de literatura verifiquei algumas características das pesquisas em MDS no Brasil. Uma delas é a de que não tem havido preferência para publicação dos trabalhos em um determinado periódico. Como a Revista Brasileira de Ciência do Solo (RBCS) é o principal meio de divulgação das pesquisas em ciência do solo do Brasil, esperava que os trabalhos estivessem ali concentrados. Contudo, dentre os dez trabalhos abordando MDS publicados no Brasil desde 2006 (ten Caten, 2011), apenas dois foram publicados na RBCS. Os demais foram publicados em outras quatro revistas nacionais. Assim, o objetivo desse questionário é identificar o motivo pelo qual as publicações envolvendo o MDS não estão concentradas na revista especializada da área e quais os principais problemas encontrados pelos autores durante o processo de submissão dos artigos. Os resultados serão utilizados em minha dissertação de mestrado e poderão ser utilizados pelos pedometristas nacionais para escolher o periódico de publicação de seus trabalhos, bem como pelos editores dos periódicos como forma de torná-los mais atrativos. Sua identificação será publicada somente em caso de concordância.

Desde já agradeço pela sua importante contribuição para o avanço do MDS no Brasil.

Questionário

1 – Quais foram os critérios utilizados para selecionar o periódico de publicação de seu trabalho?
2 – O trabalho foi negado por algum periódico? Qual foi o motivo apresentado pelo corpo editorial?
3 – No caso de ter publicado seu trabalho nas revistas Ciência Rural, Scientia Agricola, Pesquisa Agropecuária Brasileira ou Revista Brasileira de Engenharia Agrícola e Ambiental,

por que esses periódicos foram escolhidos em detrimento da Revista Brasileira de Ciência do Solo?

4 – Quais dificuldades foram encontradas durante o período de tramitação do trabalho? Você percebeu alguma forma de rejeição/conservadorismo em relação ao MDS por parte do corpo editorial e/ou revisores?

5 – A qual(is) periódico(s) você pretende submeter seus trabalhos futuros em MDS? Em que idioma? Qual o motivo?

6 – Você tem alguma sugestão a dar aos editores dos periódicos nacionais quanto à adoção de políticas de incentivo a publicação de trabalhos sobre MDS?

7 – Você concorda com a publicação de suas respostas na íntegra quando se julgar interessante? Você permite a sua identificação junto às respostas?

Informações importantes

Os trabalhos considerados nesse estudo são aqueles identificados por ten Caten (2011):

CHAGAS, C.S.; FERNANDES FILHO, E.I.; VIEIRA, C.A.O.; SCHAEFER, C.E.G.R.; CARVALHO JÚNIOR, W. Atributos topográficos e dados do Landsat7 no mapeamento digital de solos com uso de redes neurais. Pesquisa Agropecuária Brasileira, v.45, n.5, p.497-507, 2010. doi: 10.1590/S0100-204X2010000500009.

COELHO, F.F.; GIASSON, E. Comparação de métodos para mapeamento digital de solos com utilização de sistema de informação geográfica. Ciência Rural, v.40, n.10, p.2099-2106, 2010. doi: 10.1590/S0103-84782010005000156.

CRIVELENTI, R.C.; COELHO, R.M.; ADAMI, S.F.; OLIVEIRA, S.R.M. Mineração de dados para a inferência de relações solo-paisagem em mapeamentos digitais de solo. Pesquisa Agropecuária Brasileira, v.44, n.12, p.1707-1715, 2009. doi: 10.1590/S0100-204X2009001200021.

FIGUEIREDO, S.R.; GIASSON, E.; TORNQUIST, C.G.; NASCIMENTO, P.C. Uso de regressões logísticas múltiplas para mapeamento digital de solos no planalto médio do RS. Revista Brasileira de Ciência do Solo, v.32, p.2779-2785, 2008. doi: 10.1590/S0100-06832008000700023.

GIASSON, E.; SARMENTO, E.C.; WEBER, E.; FLORES, C.A.; HASENACK, H.. Decision trees for digital soil mapping on subtropical basaltic steeplands. Scientia Agrícola, v.68, p.167-174, 2011. doi: 10.1590/S0103-90162011000200006.

GIASSON, E.; CLARKE, R.T.; INDA JUNIOR, A.V.; MERTEN, G.H.; TORNQUIST, C.G. Digital soil mapping using multiple logistic regression on terrain parameters in southern Brazil. Scientia Agricola, v.63, p.262-268, 2006. doi: 10.1590/S0103-90162006000300008.

NOLASCO-CARVALHO, C.C.; FRANCA-ROCHA, W.; UCHA, J.M. Mapa digital de solos: Uma proposta metodológica usando inferência fuzzy. Revista Brasileira de Engenharia Agrícola e Ambiental, v.13, n.1, p.46-55, 2009. doi: 10.1590/S1415-43662009000100007.

TEN CATEN, A.; DALMOLIN, R.S.D.; PEDRON, F.A.; MENDONÇA-SANTOS, M.L. Regressões logísticas múltiplas: fatores que influenciam sua aplicação na predição de classes de solos. Revista Brasileira de Ciência do Solo, v.35, n.1, p.53-62, 2011a. doi: 10.1590/S0100-06832011000100005.

TEN CATEN, A.; DALMOLIN, R.S.D.; PEDRON, F.A.; MENDONÇA-SANTOS, M.L. Extrapolação das relações solo-paisagem a partir de uma área de referência. Ciência Rural, v.41, n.5, p. 812-816, 2011b. doi: 10.1590/S0103-84782011000500012

TEN CATEN, A.; DALMOLIN, R.S.D.; PEDRON, F.A.; MENDONÇA-SANTOS, M.L. Componentes principais como preditores no mapeamento digital de classes de solos. Ciência Rural, v.41, n.7, p. 1170-1176, 2011c. doi: 10.1590/S0103-84782011000700011.

Esse questionário foi enviado para o autor responsável por cada trabalho, entendido como sendo o primeiro autor ou, quanto indicado, o autor para correspondência. Em alguns casos houve sobreposição entre trabalhos. Por isso apenas seis autores foram entrevistados:

César da Silva Chagas: chagas.rj@gmail.com

Claudia C. Nolasco-Carvalho: ccseko@esser.edu.br

Elvio Giasson: giasson@ufrgs.br

Rafael Castro Crivelenti: grilasso@hotmail.com

Samuel Ribeiro Figueiredo: s.r.figueiredo@bol.com.br

Alexandre ten Caten: acaten@yahoo.com.br

Literatura citada

TEN CATEN, A. Digital soil mapping: methods to meet the demand for soil spatial information. Tese (Doutorado em Ciência do Solo) - Universidade Federal de Santa Maria, Santa Maria, RS. 2011.

b) Enviado aos editores dos periódicos nacionais onde os trabalhos foram publicados

Questionário sobre publicações abordando o mapeamento digital de solos no Brasil

Caro editor

Meu nome é Alessandro Samuel-Rosa, aluno do Programa de Pós-Graduação em Ciência do Solo da Universidade Federal de Santa Maria, sob a orientação do Prof. Dr. Ricardo Simão Diniz Dalmolin. O tema de meu projeto de pesquisa é mapeamento digital de solo (MDS). Durante a revisão de literatura verifiquei algumas características das pesquisas em MDS no Brasil. Uma delas é a de que não tem havido preferência para publicação dos trabalhos em um determinado periódico. Como a Revista Brasileira de Ciência do Solo (RBCS) é o principal meio de divulgação das pesquisas em ciência do solo do Brasil, por ser específica da área, esperava que os trabalhos estivessem ali concentrados. Contudo, dentre os dez trabalhos abordando MDS publicados no Brasil desde 2006 (ten Caten, 2011),

apenas dois foram publicados na RBCS. Os demais foram publicados em outras quatro revistas nacionais: Ciência Rural, Scientia Agricola, Pesquisa Agropecuária Brasileira ou Revista Brasileira de Engenharia Agrícola e Ambiental. Assim, o objetivo desse questionário é identificar o motivo pelo qual as publicações envolvendo o MDS não estão concentradas na revista especializada da área e quais os principais problemas encontrados pelos autores durante o processo de submissão dos artigos. Os resultados serão utilizados em minha dissertação de mestrado e poderão ser utilizados pelos pedometristas nacionais para escolher o periódico de publicação de seus trabalhos, bem como pelos editores dos periódicos como forma de torná-los mais atrativos. Sua identificação será publicada somente em caso de concordância.

Desde já agradeço pela sua importante contribuição para o avanço do MDS no Brasil.

Questionário aos editores

1 – Por que os pedometristas (pesquisadores da área da pedometria) brasileiros não concentram suas publicações na RBCS?
2 – Algum pedometrista já reclamou sobre o processo de tramitação de seus trabalhos? Quais foram as reclamações?
3 – Dentro da política adotada por seu periódico, há alguma diretriz em relação aos trabalhos que adotam o MDS?
4 – Algum revisor já apresentou posicionamento contrário ao uso das técnicas de MDS? Quais foram os argumentos?
5 – Você tem alguma sugestão a dar aos pedometristas quanto à submissão de trabalhos sobre MDS?

Informações importantes

Os trabalhos considerados nesse estudo são aqueles identificados por ten Caten (2011):

CHAGAS, C.S.; FERNANDES FILHO, E.I.; VIEIRA, C.A.O.; SCHAEFER, C.E.G.R.; CARVALHO JÚNIOR, W. Atributos topográficos e dados do Landsat7 no mapeamento digital de solos com uso de redes neurais. Pesquisa Agropecuária Brasileira, v.45, n.5, p.497-507, 2010. doi: 10.1590/S0100-204X2010000500009.

COELHO, F.F.; GIASSON, E. Comparação de métodos para mapeamento digital de solos com utilização de sistema de informação geográfica. Ciência Rural, v.40, n.10, p.2099-2106, 2010. doi: 10.1590/S0103-84782010005000156.

CRIVELENTI, R.C.; COELHO, R.M.; ADAMI, S.F.; OLIVEIRA, S.R.M. Mineração de dados para a inferência de relações solo-paisagem em mapeamentos digitais de solo. Pesquisa

Agropecuária Brasileira, v.44, n.12, p.1707-1715, 2009. doi: 10.1590/S0100-204X2009001200021.

FIGUEIREDO, S.R.; GIASSON, E.; TORNQUIST, C.G.; NASCIMENTO, P.C. Uso de regressões logísticas múltiplas para mapeamento digital de solos no planalto médio do RS. Revista Brasileira de Ciência do Solo, v.32, p.2779-2785, 2008. doi: 10.1590/S0100-06832008000700023.

GIASSON, E.; SARMENTO, E.C.; WEBER, E.; FLORES, C.A.; HASENACK, H.. Decision trees for digital soil mapping on subtropical basaltic steeplands. Scientia Agrícola, v.68, p.167-174, 2011. doi: 10.1590/S0103-90162011000200006.

GIASSON, E.; CLARKE, R.T.; INDA JUNIOR, A.V.; MERTEN, G.H.; TORNQUIST, C.G. Digital soil mapping using multiple logistic regression on terrain parameters in southern Brazil. Scientia Agricola, v.63, p.262-268, 2006. doi: 10.1590/S0103-90162006000300008.

NOLASCO-CARVALHO, C.C.; FRANCA-ROCHA, W.; UCHA, J.M. Mapa digital de solos: Uma proposta metodológica usando inferência fuzzy. Revista Brasileira de Engenharia Agrícola e Ambiental, v.13, n.1, p.46-55, 2009. doi: 10.1590/S1415-43662009000100007.

TEN CATEN, A.; DALMOLIN, R.S.D.; PEDRON, F.A.; MENDONÇA-SANTOS, M.L. Regressões logísticas múltiplas: fatores que influenciam sua aplicação na predição de classes de solos. Revista Brasileira de Ciência do Solo, v.35, n.1, p.53-62, 2011a. doi: 10.1590/S0100-06832011000100005.

TEN CATEN, A.; DALMOLIN, R.S.D.; PEDRON, F.A.; MENDONÇA-SANTOS, M.L. Extrapolação das relações solo-paisagem a partir de uma área de referência. Ciência Rural, v.41, n.5, p. 812-816, 2011b. doi: 10.1590/S0103-84782011000500012

TEN CATEN, A.; DALMOLIN, R.S.D.; PEDRON, F.A.; MENDONÇA-SANTOS, M.L. Componentes principais como preditores no mapeamento digital de classes de solos. Ciência Rural, v.41, n.7, p. 1170-1176, 2011c. doi: 10.1590/S0103-84782011000700011.

Esse questionário foi enviado para os editores dos periódicos em que os trabalhos foram publicados.

Ciência Rural: Rudi Weiblen - rudiweiblen@gmail.com

Revista Brasileira de Ciência do Solo: Roberto Ferreira Novais - rfnovais@ufv.br

Scientia Agrícola: Luís Reynaldo Ferracciú Alleoni - alleoni@esalq.usp.br

Pesquisa Agropecuária Brasileira: Emilson França de Queiroz - emilson@sede.embrapa.br

Revista Brasileira de Engenharia Agrícola e Ambiental: Hans Raj Gheyi - hans@deag.ufpb.br

Literatura citada

TEN CATEN, A. Digital soil mapping: methods to meet the demand for soil spatial information. Tese (Doutorado em Ciência do Solo) - Universidade Federal de Santa Maria, Santa Maria, RS. 2011.

Anexo 2 – Rotina das análises estatísticas realizadas no ambiente R

a) Processamento inicial dos dados utilizados nas análises

```
# Carregar um arquivo de dados para o ambiente R
dados_originais=read.table("dados_originais.csv",sep=";",head=T)

# Vincular objeto à área de trabalho
attach(dados_originais)

# Verificar os nomes das colunas do objeto fixado na área de trabalho
colnames(dados_original)

# Carregar pacote para análise estatística descritiva
library(fitdistrplus)

# Obter estatísticas descritivas das variáveis
descdist(ELEV)
descdist(EARD)
descdist(AC)
descdist(DMAC)
descdist(CONV)
descdist(CURV)
descdist(LS)
descdist(NORT)
descdist(PLAN)
descdist(PROF)
descdist(IRS)
descdist(DECL)
descdist(CD)
descdist(IPE)
descdist(IUT)

# Transformação das variáveis
AC=log(AC)
CD=sqrt(CD)
DECL=sqrt(DECL)
EARD=sqrt(EARD)
LS=sqrt(LS)
IPE=log(IPE+1)
IRS=sqrt(IRS)
IUT=log(IUT)

# Desvincular objeto da área de trabalho
detach(dados_originais)

# Definição do novo objeto contendo as variáveis transformadas
dados_final=cbind(dados_originais$clay, dados_originais$silt,
dados_originais$sand,AC,CD, dados_originais$CURV,
dados_originais$PERF,dados_originais$PLAN,DECL,dados_originais$DMAC,dados_o
riginais$ELEV,EARD,dados_originais$NORT,
LS,dados_originais$CONV,IPE,IRS,IUT)
dados=dados_final
```

```
# Obtenção dos histogramas de frequência das variáveis preditoras
attach(dados)
par(ps=22)
hist(AC,main="Histograma de AC",
ylab="Frequência",font=6,font.lab=6,font.main=6,col="gray")
hist(CD,main="Histograma de CD",
ylab="Frequência",font=6,font.lab=6,font.main=6,col="gray")
hist(CONV,main="Histograma de CONV",
ylab="Frequência",font=6,font.lab=6,font.main=6,col="gray")
hist(CURV,main="Histograma de CURV",
ylab="Frequência",font=6,font.lab=6,font.main=6,col="gray")
hist(DECL,main="Histograma de DECL",
ylab="Frequência",font=6,font.lab=6,font.main=6,col="gray")
hist(DMAC,main="Histograma de DMAC",
ylab="Frequência",font=6,font.lab=6,font.main=6,col="gray")
hist(EARD,main="Histograma de EARD",
ylab="Frequência",font=6,font.lab=6,font.main=6,col="gray")
hist(ELEV,main="Histograma de ELEV",
ylab="Frequência",font=6,font.lab=6,font.main=6,col="gray")
hist(NORT,main="Histograma de NORT",
ylab="Frequência",font=6,font.lab=6,font.main=6,col="gray")
hist(PLAN,main="Histograma de PLAN",
ylab="Frequência",font=6,font.lab=6,font.main=6,col="gray")
hist(PERF,main="Histograma de PERF",
ylab="Frequência",font=6,font.lab=6,font.main=6,col="gray")
hist(LS,main="Histograma de LS",
ylab="Frequência",font=6,font.lab=6,font.main=6,col="gray")
hist(IRS,main="Histograma de IRS",
ylab="Frequência",font=6,font.lab=6,font.main=6,col="gray")
hist(IPE,main="Histograma de IPE",
ylab="Frequência",font=6,font.lab=6,font.main=6,col="gray")
hist(IUT,main="Histograma de IUT",
ylab="Frequência",font=6,font.lab=6,font.main=6,col="gray")
detach(dados)

# Particionar o conjunto de dados completo (n = 339) nos dois domínios
fisiográficos
dados_inf=dados[which(dados$ELEV<300),]
dados_sup=dados[which(dados$ELEV>=300),]

# Obter amostras aleatórias
set.seed(123)
sample1=dados[sample(nrow(dados),300),]
set.seed(123)
sample.inf=dados_inf[sample(nrow(dados_inf),150),]
set.seed(123)
sample.sup=dados_sup[sample(nrow(dados_sup),150),]

# Elaboração do triângulo textural
library(soiltexture)
attach(dados)
texture=cbind(parent,clay,silt,sand)
detach(dados)
```

```
TT.plot(tri.data=texture,css.lab=c("Argila (%)","Silte (%)","Areia
(%)"),font.axis=6,font.lab=6,grid.col="gray",class.sys="none",frame.bg.col=
"white", cex=1,lwd.axis=2,lwd.lab=2,col=texture[,"parent"],pch=20)

# Elaboração do gráfico de tendência do MDS no Brasil e no mundo
mds=read.table("mds.txt",sep=",",head=T)
attach(mds)
plot(ano,mundo/10,xlab="Ano",ylab="N° de
ocorrências",xaxp=c(2003,2011,8),font.lab=6,font.axis=6,pch=20,cex=2)
points(ano,brasil,pch=1,cex=2)
lines(lowess(ano,mundo/10),lwd=2)
lines(lowess(ano,brasil),lwd=2,lty=2)
legend(2003,25,legenda,pch=c(1,20),lty=c(2,1),cex=1.2)
```

b) Análise de componentes principais

```
attach(dados)

# Definição dos dados a serem utilizados na análise de componentes
principais
dados_pca=cbind(AC,CD,CURV,DECL,DMAC,ELEV,EARD,NORT,CONV,IPE,IUT)
detach(dados)

# Obtenção da matriz de correlação
R_dados_pca=cor(dados_pca)

# Carregar o pacote necessário para o teste de esfericidade de Bartlett
library(psych)

# Realização do teste de esfericidade de Bartlett
cortest.bartlett(R_dados_pca,339)

# Código-fonte dos testes de adequação amostral KMO (Kaiser-Meyer-Olkin) e
MSA (Measure of Sample Adequacy)
kmo = function( data ){

   library(MASS)
   X <- cor(as.matrix(data))
   iX <- ginv(X)
   S2 <- diag(diag((iX^-1)))
   AIS <- S2%*%iX%*%S2                      # anti-image covariance matrix
   IS <- X+AIS-2*S2                         # image covariance matrix
   Dai <- sqrt(diag(diag(AIS)))
   IR <- ginv(Dai)%*%IS%*%ginv(Dai)         # image correlation matrix
   AIR <- ginv(Dai)%*%AIS%*%ginv(Dai)       # anti-image correlation matrix
   a <- apply((AIR - diag(diag(AIR)))^2, 2, sum)
   AA <- sum(a)
   b <- apply((X - diag(nrow(X)))^2, 2, sum)
   BB <- sum(b)
   MSA <- b/(b+a)                    # indiv. measures of sampling adequacy

   AIR <- AIR-diag(nrow(AIR))+diag(MSA)  # Examine the anti-image of the
                                         # correlation matrix. That is the
                                    # negative of the partial correlations,
                                    # partialling out all other variables.
```

```
   kmo <- BB/(AA+BB)                    # overall KMO statistic

   # Reporting the conclusion
     if (kmo >= 0.00 && kmo < 0.50){
       test <- 'The KMO test yields a degree of common variance
unacceptable for FA.'
     } else if (kmo >= 0.50 && kmo < 0.60){
       test <- 'The KMO test yields a degree of common variance miserable.'
     } else if (kmo >= 0.60 && kmo < 0.70){
       test <- 'The KMO test yields a degree of common variance mediocre.'
     } else if (kmo >= 0.70 && kmo < 0.80){
       test <- 'The KMO test yields a degree of common variance middling.'
     } else if (kmo >= 0.80 && kmo < 0.90){
       test <- 'The KMO test yields a degree of common variance
meritorious.'
     } else {
       test <- 'The KMO test yields a degree of common variance marvelous.'
     }

     ans <- list(  overall = kmo,
                   report = test,
                   individual = MSA,
                   AIS = AIS,
                   AIR = AIR )
     return(ans)

}    # end of kmo()

# Remover a variável NORTH do conjunto de dados
dados_pca$NORTH=NULL

# Calcular autovalores e autovetores da matriz de correlação
auto1=eigen(cor(dados_pca))

# Gerar scree-plot
plot(auto1$values,type="l",xlab="Componente
principal",ylab="Autovalor",font.lab=6,font.axis=6)
points(auto1$values,pch=20)
abline(h=1)

# Realizar a análise de componentes principais
pca1=princomp(dados_pca,cor=T)
summary(pca1)
print(pca1$loadings,digits=3,cutoff=NULL)
pca1$scores

# Calcular a correlação entre as variáveis originais e os escores das
componentes principais
CP1=pca1$loadings[,1]*sqrt(auto1$values[1])
CP2=pca1$loadings[,2]*sqrt(auto1$values[2])
CP3=pca1$loadings[,3]*sqrt(auto1$values[3])
```

```
# Gerar os gráficos das projeções
library(plotrix)
par(ps=16)
proj1=cbind(CP1,CP2)
proj2=cbind(CP2,CP3)
plot(proj1,asp=1,xlim=c(-1,1),ylim=c(-1,1),xlab="CP 1 (41%)",ylab="CP 2
(26%)",pch=20,font.lab=6,font.axis=6)
draw.circle(0,0,1)
abline(v=0,h=0)
text(proj1,rownames(proj1),pos=2,font=6)
plot(proj2,asp=1,xlim=c(-1,1),ylim=c(-1,1),xlab="CP 2 (26%)",ylab="CP 3
(11%)",pch=20,font.lab=6,font.axis=6)
draw.circle(0,0,1)
abline(v=0,h=0)
text(proj2,rownames(proj2),pos=1,font=6)
```

c) Análise de regressão linear múltipla e obtenção de estatísticas para sua avaliação

```
attach(sample1)
library(MASS)
library(HH)

# Análise de regressão da variável ln(argila/areia) para toda a área de
estudo
fit.clay.sand=lm(log(clay/sand)~CONV+ELEV+IPE+IUT)

# Seleção das variáveis através do método stepwise
fit.clay.sand=stepAIC(fit.clay.sand,scope=list(upper=~ELEV+IPE+CONV+IUT,low
er=~1),direction="both",trace=FALSE)

# Obtenção das estísticas da análise de regressão
summary(fit.clay.sand)
anova(fit.clay.sand)
extractAIC(fit.clay.sand)
vcov(fit.clay.sand)
vif(fit.clay.sand)

# Geração dos gráficos para análise dos resíduos
par(ps=18)
par(font=6)
plot(fit.clay.sand,which=1,font=6,font.lab=6,font.axis=6,font.main=6,pch=20
,font.sub=6,lwd=3,id.n=6, labels.id=sample1$id)
plot(fit.clay.sand,which=2,font=6,font.lab=6,font.axis=6,font.main=6,pch=20
,font.sub=6,lwd=3,id.n=6, labels.id=sample1$id)
plot(fit.clay.sand,which=4,font=6,font.lab=6,font.axis=6,font.main=6,pch=20
,font.sub=6,lwd=3,id.n=6, labels.id=sample1$id)
plot(fit.clay.sand,which=5,font=6,font.lab=6,font.axis=6,font.main=6,pch=20
,font.sub=6,lwd=3,id.n=6, labels.id=sample1$id)

# Análise de regressão da variável ln(silte/areia) para toda a área de
estudo
fit.silt.sand=lm(log(silt/sand)~CONV+ELEV+IPE+IUT)
```

```
fit.silt.sand=stepAIC(fit.silt.sand,scope=list(upper=~ELEV+IPE+CONV+IUT,low
er=~1),direction="both”)
summary(fit.silt.sand)
anova(fit.silt.sand)
extractAIC(fit.silt.sand)
vcov(fit.silt.sand)
vif(fit.silt.sand)
plot(fit.silt.sand,which=1,font=6,font.lab=6,font.axis=6,font.main=6,pch=20
,font.sub=6,lwd=3,id.n=6, labels.id=sample1$id)
plot(fit.silt.sand,which=2,font=6,font.lab=6,font.axis=6,font.main=6,pch=20
,font.sub=6,lwd=3,id.n=6, labels.id=sample1$id)
plot(fit.silt.sand,which=4,font=6,font.lab=6,font.axis=6,font.main=6,pch=20
,font.sub=6,lwd=3,id.n=6, labels.id=sample1$id)
plot(fit.silt.sand,which=5,font=6,font.lab=6,font.axis=6,font.main=6,pch=20
,font.sub=6,lwd=3,id.n=6, labels.id=sample1$id)
detach(sample1)

# Análise de regressão da variável ln(argila/areia) para o domínio
fisiográfico inferior
attach(sample.inf)
fit.clay.sand.inf=lm(log(clay/sand)~CONV+ELEV+IPE+IUT)
fit.clay.sand.inf=stepAIC(fit.clay.sand.inf,scope=list(upper=~ELEV+IPE+CONV
+IUT,lower=~1),direction="both”)
summary(fit.clay.sand.inf)
anova(fit.clay.sand.inf)
extractAIC(fit.clay.sand.inf)
vcov(fit.clay.sand.inf)
vif(fit.clay.sand.inf)
par(ps=18)
par(font=6)
plot(fit.clay.sand.inf,which=1,font=6,font.lab=6,font.axis=6,font.main=6,pc
h=20,font.sub=6,lwd=3,id.n=6, labels.id=sample.inf$id)
plot(fit.clay.sand.inf,which=2,font=6,font.lab=6,font.axis=6,font.main=6,pc
h=20,font.sub=6,lwd=3,id.n=6, labels.id=sample.inf$id)
plot(fit.clay.sand.inf,which=4,font=6,font.lab=6,font.axis=6,font.main=6,pc
h=20,font.sub=6,lwd=3,id.n=6, labels.id=sample.inf$id)
plot(fit.clay.sand.inf,which=5,font=6,font.lab=6,font.axis=6,font.main=6,pc
h=20,font.sub=6,lwd=3,id.n=6, labels.id=sample.inf$id)

# Análise de regressão da variável ln(silte/areia) para o domínio
fisiográfico inferior
fit.silt.sand.inf=lm(log(silt/sand)~CONV+ELEV+IPE+IUT)
fit.silt.sand.inf=stepAIC(fit.silt.sand.inf,scope=list(upper=~ELEV+IPE+CONV
+IUT,lower=~1),direction="both")
anova(fit.silt.sand.inf)
summary(fit.silt.sand.inf)
extractAIC(fit.silt.sand.inf)
vcov(fit.silt.sand.inf)
vif(fit.silt.sand.inf)
plot(fit.silt.sand.inf,which=1,font=6,font.lab=6,font.axis=6,font.main=6,pc
h=20,font.sub=6,lwd=3,id.n=6, labels.id=sample.inf$id)
plot(fit.silt.sand.inf,which=2,font=6,font.lab=6,font.axis=6,font.main=6,pc
h=20,font.sub=6,lwd=3,id.n=6, labels.id=sample.inf$id)
```

```
plot(fit.silt.sand.inf,which=4,font=6,font.lab=6,font.axis=6,font.main=6,pc
h=20,font.sub=6,lwd=3,id.n=6, labels.id=sample.inf$id)
plot(fit.silt.sand.inf,which=5,font=6,font.lab=6,font.axis=6,font.main=6,pc
h=20,font.sub=6,lwd=3,id.n=6, labels.id=sample.inf$id)
detach(sample.inf)

# Análise de regressão da variável ln(argila/areia) para o domínio
fisiográfico superior
attach(sample.sup)
fit.clay.sand.sup=lm(log(clay/sand)~CONV+ELEV+IPE+IUT)
fit.clay.sand.sup=stepAIC(fit.clay.sand.sup,scope=list(upper=~CONV+ELEV+IPE
+IUT,lower=~1),direction="both")
anova(fit.clay.sand.sup)
summary(fit.clay.sand.sup)
extractAIC(fit.clay.sand.sup)
vcov(fit.clay.sand.sup)
vif(fit.clay.sand.sup)
plot(fit.clay.sand.sup,which=1,font=6,font.lab=6,font.axis=6,font.main=6,pc
h=20,font.sub=6,lwd=3,id.n=6, labels.id=sample.sup$id)
plot(fit.clay.sand.sup,which=2,font=6,font.lab=6,font.axis=6,font.main=6,pc
h=20,font.sub=6,lwd=3,id.n=6, labels.id=sample.sup$id)
plot(fit.clay.sand.sup,which=4,font=6,font.lab=6,font.axis=6,font.main=6,pc
h=20,font.sub=6,lwd=3,id.n=6, labels.id=sample.sup$id)
plot(fit.clay.sand.sup,which=5,font=6,font.lab=6,font.axis=6,font.main=6,pc
h=20,font.sub=6,lwd=3,id.n=6, labels.id=sample.sup$id)

# Análise de regressão da variável ln(silte/areia) para o domínio
fisiográfico superior
fit.silt.sand.sup=lm(log(silt/sand)~CONV+ELEV+IPE+IUT)
fit.silt.sand.sup=stepAIC(fit.silt.sand.sup,scope=list(upper=~CONV+ELEV+IPE
+IUT,lower=~1),direction="both")
anova(fit.silt.sand.sup)
summary(fit.silt.sand.sup)
extractAIC(fit.silt.sand.sup)
vcov(fit.silt.sand.sup)
vif(fit.silt.sand.sup)
plot(fit.silt.sand.sup,which=1,font=6,font.lab=6,font.axis=6,font.main=6,pc
h=20,font.sub=6,lwd=3,id.n=6, labels.id=sample.sup$id)
plot(fit.silt.sand.sup,which=2,font=6,font.lab=6,font.axis=6,font.main=6,pc
h=20,font.sub=6,lwd=3,id.n=6, labels.id=sample.sup$id)
plot(fit.silt.sand.sup,which=4,font=6,font.lab=6,font.axis=6,font.main=6,pc
h=20,font.sub=6,lwd=3,id.n=6, labels.id=sample.sup$id)
plot(fit.silt.sand.sup,which=5,font=6,font.lab=6,font.axis=6,font.main=6,pc
h=20,font.sub=6,lwd=3,id.n=6, labels.id=sample.sup$id)
detach(sample.sup)

# Transformação dos valores preditos (log-razões aditivas) para as frações
de tamanho de partícula
attach(sample1)
sand.pred=(1/(exp(fit.clay.sand$fitted)+exp(fit.silt.sand$fitted)+1))*100
silt.pred=(exp(fit.silt.sand$fitted)/(exp(fit.clay.sand$fitted)+exp(fit.sil
t.sand$fitted)+1))*100
```

```
clay.pred=(exp(fit.clay.sand$fitted)/(exp(fit.clay.sand$fitted)+exp(fit.sil
t.sandfitted)+1)) *100

# Obtenção dos gráficos de comparação dos valores preditos e medidos de
cada fração
plot(sand,pred.sand,font=6,pch=20,xlab="Areia medida (%)",ylab="Areia
predita (%)", ylim=c(0,100),xlim=c(0,100), font.lab=6,cex=1.5)
abline(a=0,b=1)
plot(silt,pred.silt,font=6,pch=20,xlab="Silte medido (%)",ylab="Silte
predito (%)",ylim=c(0,100),xlim=c(0,100),
font.lab=6,cex=1.5)
abline(a=0,b=1)
plot(clay,pred.clay,font=6,pch=20,xlab="Argila medida (%)",ylab="Argila
predita (%)",ylim=c(0,100),xlim=c(0,100),
font.lab=6,cex=1.5)
abline(a=0,b=1)
detach(sample1)

# Transformação dos valores preditos (log-razões aditivas) para as frações
de tamanho de partícula
attach(sample_inf)
AREIA.PRED.INF=(1/(exp(fit.clay.sand.inf$fitted)+exp(fit.silt.sand.inf$fitt
ed)+1))*100
SILTE.PRED.INF=(exp(fit.silt.sand.inf$fitted)/(exp(fit.clay.sand.inf$fitted
)+exp(fit.silt.sand.inf$fitted)+1))*100
ARGILA.PRED.INF=(exp(fit.clay.sand.inf$fitted)/(exp(fit.clay.sand.inf$fitte
d)+exp(fit.silt.sand.inf$fitted)+1)) *100

# Obtenção dos gráficos de comparação dos valores preditos e medidos de
cada fração
plot(sand,AREIA.PRED.INF,font=6,pch=20,xlab="Areia medida (%)",ylab="Areia
predita (%)",ylim=c(0,100), xlim=c(0,100),font.lab=6,cex=1.5)
abline(a=0,b=1)
plot(silt,SILTE.PRED.INF,font=6,pch=20,xlab="Silte medido (%)",ylab="Silte
predito (%)",ylim=c(0,100), xlim=c(0,100),font.lab=6,cex=1.5)
abline(a=0,b=1)
plot(clay,ARGILA.PRED.INF,font=6,pch=20,xlab="Argila medida
(%)",ylab="Argila medida (%)",ylim=c(0,100),
xlim=c(0,100),font.lab=6,cex=1.5)
abline(a=0,b=1)
detach(sample.inf)

# Transformação dos valores preditos (log-razões aditivas) para as frações
de tamanho de partícula
attach(sample.sup)
AREIA.PRED.SUP=(1/(exp(fit.clay.sand.sup$fitted)+exp(fit.silt.sand.sup$fitt
ed)+1))*100
ARGILA.PRED.SUP=(exp(fit.clay.sand.sup$fitted)/(exp(fit.clay.sand.sup$fitte
d)+exp(fit.silt.sand.sup$fitted)+1)) *100
SILTE.PRED.SUP=(exp(fit.silt.sand.sup$fitted)/(exp(fit.clay.sand.sup$fitted
)+exp(fit.silt.sand.sup$fitted)+1))*100
```

```
# Obtenção dos gráficos de comparação dos valores preditos e medidos de
cada fração
plot(sand,AREIA.PRED.SUP,font=6,pch=20,xlab="Areia medida (%)",ylab="Areia
predita (%)", ylim=c(0,100),xlim=c(0,100),font.lab=6,cex=1.5)
abline(a=0,b=1)
plot(silt,SILTE.PRED.SUP,font=6,pch=20,xlab="Silte medido (%)",ylab="Silte
predito (%)", ylim=c(0,100),xlim=c(0,100),font.lab=6,cex=1.5)
abline(a=0,b=1)
plot(clay,ARGILA.PRED.SUP,font=6,pch=20,xlab="Argila medida
(%)",ylab="Argila medida (%)",
ylim=c(0,100),xlim=c(0,100),font.lab=6,cex=1.5)
abline(a=0,b=1)
detach(sample.sup)

# Predição da distribuição do tamanho de partícula das 339 observações e
cálculo dos resíduos p/ análise geoestatística
attach(dados)
pred.argila.areia=predict(fit.clay.sand,dados)
pred.silte.areia=predict(fit.silt.sand,dados)
pred.argila=((exp(pred.argila.areia))/(exp(pred.argila.areia)+exp(pred.silt
e.areia)+1))*100
pred.silte=((exp(pred.silte.areia))/(exp(pred.argila.areia)+exp(pred.silte.
areia)+1))*100
pred.areia=(1/(exp(pred.argila.areia)+exp(pred.silte.areia)+1))*100
res.argila=pred.argila-clay
res.silte=pred.silte-silt
res.areia=pred.areia-sand
residuos=cbind(res.argila,res.silte,res.areia)

# Salvar arquivo csv contendo os resíduos
write.csv(residuos,"residuos")

# Identificação das observações atípicas e influenciais
erro1=rbind(sample1[which(id==328),],sample1[which(id==162),],sample1[which
(id==121),],sample1[which(id==99),],sample1[which(id==116),],sample1[which(
id==286),])
infl1=rbind(sample1[which(id==172),],sample1[which(id==328),],sample1[which
(id==162),],sample1[which(id==140),],sample1[which(id==160),],sample1[which
(id==302),])
erro2=rbind(sample1[which(id==161),],sample1[which(id==286),],sample1[which
(id==101),],sample1[which(id==99),],sample1[which(id==162),],sample1[which(
id==135),])
infl2=rbind(sample1[which(id==1),],sample1[which(id==162),],sample1[which(i
d==140),],sample1[which(id==166),],sample1[which(id==173),],sample1[which(i
d==302),])
erro3=rbind(sample1[which(id==1),],sample1[which(id==39),],sample1[which(id
==162),],sample1[which(id==135),],sample1[which(id==121),],sample1[which(id
==45),])
infl3=rbind(sample1[which(id==162),],sample1[which(id==160),],sample1[which
(id==302),],sample1[which(id==1),],sample1[which(id==121),],sample1[which(i
d==135),])
```

```
erro4=rbind(sample1[which(id==162),],sample1[which(id==161),],sample1[which
(id==1),],sample1[which(id==68),],sample1[which(id==135),],sample1[which(id
==121),])
infl4=rbind(sample1[which(id==160),],sample1[which(id==162),],sample1[which
(id==302),],sample1[which(id==1),],sample1[which(id==121),],sample1[which(i
d==135),])
erro5=rbind(sample1[which(id==328),],sample1[which(id==116),],sample1[which
(id==117),],sample1[which(id==99),],sample1[which(id==101),],sample1[which(
id==226),])
infl5=rbind(sample1[which(id==116),],sample1[which(id==289),],sample1[which
(id==142),],sample1[which(id==328),],sample1[which(id==237),],sample1[which
(id==117),])
erro6=rbind(sample1[which(id==116),],sample1[which(id==117),],sample1[which
(id==101),],sample1[which(id==99),],sample1[which(id==98),],sample1[which(i
d==103),])
infl6=rbind(sample1[which(id==116),],sample1[which(id==117),],sample1[which
(id==99),],sample1[which(id==101),],sample1[which(id==311),],sample1[which(
id==227),])

# Salvar arquivos csv contendo os dados
write.csv(erro1,"erro1")
write.csv(erro2,"erro2")
write.csv(erro3,"erro3")
write.csv(erro4,"erro4")
write.csv(erro5,"erro5")
write.csv(erro6,"erro6")
write.csv(infl1,"infl1")
write.csv(infl2,"infl2")
write.csv(infl3,"infl3")
write.csv(infl4,"infl4")
write.csv(infl5,"infl5")
write.csv(infl6,"infl6")
```

d) Validação cruzada e cálculo das estatísticas descritivas

```
# carregar os pacotes necessários
library(caret)
library(timeSeries)

# obter a identificação das observações usadas como conjunto de teste na
validação cruzada
obs.id <- function(y, k = 10, times = 5)
   {
     prettyNums <- paste("Rep", gsub(" ", "0", format(1:times)), sep = "")
     for(i in 1:times)
       {
         tmp <- createFolds(y, k = k, list = TRUE, returnTrain = FALSE)
 names(tmp) <- paste("Fold",
                             gsub(" ", "0", format(seq(along = tmp))),
                             ".",
                             prettyNums[i],
                             sep = "")
         out <- if(i == 1) tmp else c(out, tmp)
```

```
      }
    out
  }

# código-fonte para calcular o coeficiente de correlação ajustado
r2aj=function(r2,n,p){r2aj=1-(1-r2)*((n-1)/(n-p-
1));ans=list(r2aj);return(ans)}

# Validação cruzada das FPESe construídas para toda a área de estudo
# Definição dos conjuntos de dados de variáveis dependentes e preditoras
attach(dados)
preditoras=cbind(CONV,ELEV,IPE)
ln.clay.sand=log(clay/sand)
ln.silt.sand=log(silt/sand)
detach(dados)

# validação cruzada da FPESe para estimar ln(clay/sand) em toda a área de
estudo
set.seed(123)
tmp.clay.sand=createMultiFolds(ln.clay.sand,k=10,times=100)
trControl.clay.sand=trainControl(method="repeatedcv",number=10,repeats=100,
savePredictions=TRUE,index=tmp.clay.sand)
cv.clay.sand=train(preditoras,ln.clay.sand,"lm",trControl=trControl.clay.sa
nd)

# identificação das observações usadas como conjunto de teste
set.seed(123)
cv.clay.sand.ID=obs.id(ln.clay.sand,k=10,times=100)
L.max <- max(sapply(cv.clay.sand.ID, length), na.rm = TRUE)
cv.clay.sand.ID =t(sapply(cv.clay.sand.ID, function(x) c(x, rep(NA, L.max -
length(x)))))
cv.clay.sand.ID =as.numeric(na.omit(c(t(as.data.frame(cv.clay.sand.ID)))))

# Organização dos dados
cv.clay.sand=as.data.frame(cv.clay.sand$pred[1:2])
cv.rep=sort(rep(1:100,339))
cv.clay.sand=cbind(cv.rep,cv.clay.sand.ID,cv.clay.sand)
rm(cv.clay.sand.ID, cv.rep)
attach(cv.clay.sand)
cv.clay.sand=cv.clay.sand[order(cv.rep,cv.clay.sand.ID),]
detach(cv.clay.sand)

# validação cruzada da FPESe para estimar ln(silt/sand) em toda a área de
estudo
set.seed(123)
tmp.silt.sand=createMultiFolds(ln.silt.sand,k=10,times=100)
trControl.silt.sand=trainControl(method="repeatedcv",number=10,repeats=100,
savePredictions=TRUE,index=tmp.silt.sand)
cv.silt.sand=train(preditoras,ln.silt.sand,"lm",trControl=trControl.silt.sa
nd)
```

```
# identificação das observações usadas como conjunto de teste
set.seed(123)
cv.silt.sand.ID=obs.id(ln.silt.sand,k=10,times=100)
L.max <- max(sapply(cv.silt.sand.ID, length), na.rm = TRUE)
cv.silt.sand.ID =t(sapply(cv.silt.sand.ID, function(x) c(x, rep(NA, L.max -
length(x)))))
cv.silt.sand.ID =as.numeric(na.omit(c(t(as.data.frame(cv.silt.sand.ID)))))

# Organização dos dados
cv.silt.sand=as.data.frame(cv.silt.sand$pred[1:2])
cv.rep=sort(rep(1:100,339))
cv.silt.sand=cbind(cv.rep,cv.silt.sand.ID,cv.silt.sand)
rm(cv.silt.sand.ID, cv.rep)
attach(cv.silt.sand)
cv.silt.sand=cv.silt.sand[order(cv.rep,cv.silt.sand.ID),]
detach(cv.silt.sand)

# Organização dos dados
sand.obs=(         1
/(exp(cv.clay.sand$obs)+exp(cv.silt.sand$obs)+1))*100
silt.obs=((exp(cv.silt.sand$obs))/(exp(cv.clay.sand$obs)+exp(cv.silt.sand$o
bs)+1))*100
clay.obs=((exp(cv.clay.sand$obs))/(exp(cv.clay.sand$obs)+exp(cv.silt.sand$o
bs)+1))*100
sand.obs=matrix(sand.obs,nrow=339)
silt.obs=matrix(silt.obs,nrow=339)
clay.obs=matrix(clay.obs,nrow=339)
sand.pred=(1/(exp(cv.clay.sand$pred)+exp(cv.silt.sand$pred)+1))*100
silt.pred=(exp(cv.silt.sand$pred)/(exp(cv.clay.sand$pred)+exp(cv.silt.sand$
pred)+1))*100
clay.pred=(exp(cv.clay.sand$pred)/(exp(cv.clay.sand$pred)+exp(cv.silt.sand$
pred)+1))*100
sand.pred=matrix(sand.pred,nrow=339)
silt.pred=matrix(silt.pred,nrow=339)
clay.pred=matrix(clay.pred,nrow=339)

# Cálculo das estatísticas da predição da areia
sand.res=sand.pred-sand.obs
sand.me=colMeans(sand.res)
sand.eqm=colMeans(sand.res^2)
sand.rmse=sqrt(sand.eqm)
sand.r2=colSums((sand.pred-colMeans(sand.obs))^2)/colSums((sand.obs-
colMeans(sand.obs))^2)
sand.r2aj=as.numeric((r2aj(sand.r2,300,4))[[1]])
sand.int=(colMaxs(sand.obs)-colMins(sand.obs))
sand.rmse.n=sand.rmse/sand.int
sand.stats=cbind(sand.rmse,sand.rmse.n,sand.me,sand.r2aj)
mediana=matrix(c(apply(sand.stats,2,median)),nrow=4)
percentis=t(matrix(c(quantile(sand.rmse,c(0.025,0.975)),quantile(sand.rmse.
n,c(0.025,0.975)),quantile(sand.me,c(0.025,0.975)),
quantile(sand.r2aj,c(0.025,0.975))),nrow=2))
```

```
sand.stats=cbind(percentis[,1],mediana,percentis[,2]);colnames(sand.stats)=
c("2,5 percentil","Mediana","97,5
percentil");rownames(sand.stats)=c("RMSE","RMSEnorm","EM","R2aj")

# Cálculo das estatísticas da predição do silte
silt.res=silt.pred-silt.obs
silt.me=colMeans(silt.res)
silt.eqm=colMeans(silt.res^2)
silt.rmse=sqrt(silt.eqm)
silt.r2=colSums((silt.pred-colMeans(silt.obs))^2)/colSums((silt.obs-
colMeans(silt.obs))^2)
silt.r2aj=as.numeric((r2aj(silt.r2,300,4))[[1]])
silt.int=(colMaxs(silt.obs)-colMins(silt.obs))
silt.rmse.n=silt.rmse/silt.int
silt.stats=cbind(silt.rmse,silt.rmse.n,silt.me,silt.r2aj)
mediana=matrix(c(apply(silt.stats,2,median)),nrow=4)
percentis=t(matrix(c(quantile(silt.rmse,c(0.025,0.975)),quantile(silt.rmse.
n,c(0.025,0.975)),quantile(silt.me,c(0.025,0.975)),
quantile(silt.r2aj,c(0.025,0.975))),nrow=2))
silt.stats=cbind(percentis[,1],mediana,percentis[,2]);colnames(silt.stats)=
c("2,5 percentil","Mediana","97,5
percentil");rownames(silt.stats)=c("RMSE","RMSEnorm","EM","R2aj")

# Cálculo das estatísticas da predição da argila
clay.res=clay.pred-clay.obs
clay.me=colMeans(clay.res)
clay.eqm=colMeans(clay.res^2)
clay.rmse=sqrt(clay.eqm)
clay.r2=colSums((clay.pred-colMeans(clay.obs))^2)/colSums((clay.obs-
colMeans(clay.obs))^2)
clay.r2aj=as.numeric((r2aj(clay.r2,300,4))[[1]])
clay.int=(colMaxs(clay.obs)-colMins(clay.obs))
clay.rmse.n=clay.rmse/clay.int
clay.stats=cbind(clay.rmse,clay.rmse.n,clay.me,clay.r2aj)
mediana=matrix(c(apply(clay.stats,2,median)),nrow=4)
percentis=t(matrix(c(quantile(clay.rmse,c(0.025,0.975)),quantile(clay.rmse.
n,c(0.025,0.975)),quantile(clay.me,c(0.025,0.975)),
quantile(clay.r2aj,c(0.025,0.975))),nrow=2))
clay.stats=cbind(percentis[,1],mediana,percentis[,2]);colnames(clay.stats)=
c("2,5 percentil","Mediana","97,5
percentil");rownames(clay.stats)=c("RMSE","RMSEnorm","EM","R2aj")

# Estatísticas da validação cruzada
sand.stats
silt.stats
Clay.stats

# Definição das variáveis dependentes e preditoras no domínio fisiográfico
inferior
attach(dados_inf)
preditoras=as.data.frame(cbind(CONV,ELEV,IPE))
ln.clay.sand=log(clay/sand)
ln.silt.sand=log(silt/sand)
```

```
detach(dados_inf)


# validação cruzada da FPESe para estimar ln(clay/sand)
set.seed(123)
tmp.clay.sand=createMultiFolds(ln.clay.sand,k=10,times=100)
trControl.clay.sand=trainControl(method="repeatedcv",number=10,repeats=100,
savePredictions=TRUE,index=tmp.clay.sand)
cv.clay.sand=train(preditoras,ln.clay.sand,"lm",trControl=trControl.clay.sa
nd)

# identificação das observações usadas como conjunto de teste
set.seed(123)
cv.clay.sand.ID=obs.id(ln.clay.sand,k=10,times=100)
L.max <- max(sapply(cv.clay.sand.ID, length), na.rm = TRUE)
cv.clay.sand.ID =t(sapply(cv.clay.sand.ID, function(x) c(x, rep(NA, L.max -
length(x)))))
cv.clay.sand.ID =as.numeric(na.omit(c(t(as.data.frame(cv.clay.sand.ID)))))

# organização dos dados para cálculo das estatísticas da validação cruzada
cv.clay.sand=as.data.frame(cv.clay.sand$pred[1:2])
cv.rep=sort(rep(1:100,165))
cv.clay.sand=cbind(cv.rep,cv.clay.sand.ID,cv.clay.sand)
rm(cv.clay.sand.ID, cv.rep)
attach(cv.clay.sand)
cv.clay.sand=cv.clay.sand[order(cv.rep,cv.clay.sand.ID),]
detach(cv.clay.sand)

# validação cruzada da FPESe para estimar ln(silt/sand)
set.seed(123)
tmp.silt.sand=createMultiFolds(ln.silt.sand,k=10,times=100)
trControl.silt.sand=trainControl(method="repeatedcv",number=10,repeats=100,
savePredictions=TRUE,index=tmp.silt.sand)
cv.silt.sand=train(preditoras,ln.silt.sand,"lm",trControl=trControl.silt.sa
nd)

# identificação das observações usadas como conjunto de teste
set.seed(123)
cv.silt.sand.ID=obs.id(ln.silt.sand,k=10,times=100)
L.max <- max(sapply(cv.silt.sand.ID, length), na.rm = TRUE)
cv.silt.sand.ID =t(sapply(cv.silt.sand.ID, function(x) c(x, rep(NA, L.max -
length(x)))))
cv.silt.sand.ID =as.numeric(na.omit(c(t(as.data.frame(cv.silt.sand.ID)))))

# organização dos dados para cálculo das estatísticas da validação cruzada
cv.silt.sand=as.data.frame(cv.silt.sand$pred[1:2])
cv.rep=sort(rep(1:100,165))
cv.silt.sand=cbind(cv.rep,cv.silt.sand.ID,cv.silt.sand)
rm(cv.silt.sand.ID, cv.rep)
attach(cv.silt.sand)
cv.silt.sand=cv.silt.sand[order(cv.rep,cv.silt.sand.ID),]
detach(cv.silt.sand)
```

```
# Organização dos dados
sand.obs=(         1
/(exp(cv.clay.sand$obs)+exp(cv.silt.sand$obs)+1))*100
silt.obs=((exp(cv.silt.sand$obs))/(exp(cv.clay.sand$obs)+exp(cv.silt.sand$o
bs)+1))*100
clay.obs=((exp(cv.clay.sand$obs))/(exp(cv.clay.sand$obs)+exp(cv.silt.sand$o
bs)+1))*100
sand.obs=matrix(sand.obs,nrow=165)
silt.obs=matrix(silt.obs,nrow=165)
clay.obs=matrix(clay.obs,nrow=165)
sand.pred=(1/(exp(cv.clay.sand$pred)+exp(cv.silt.sand$pred)+1))*100
silt.pred=(exp(cv.silt.sand$pred)/(exp(cv.clay.sand$pred)+exp(cv.silt.sand$
pred)+1))*100
clay.pred=(exp(cv.clay.sand$pred)/(exp(cv.clay.sand$pred)+exp(cv.silt.sand$
pred)+1))*100
sand.pred=matrix(sand.pred,nrow=165)
silt.pred=matrix(silt.pred,nrow=165)
clay.pred=matrix(clay.pred,nrow=165)

# Cálculo das estatísticas da predição da areia
sand.res=sand.pred-sand.obs
sand.me=colMeans(sand.res)
sand.eqm=colMeans(sand.res^2)
sand.rmse=sqrt(sand.eqm)
sand.r2=colSums((sand.pred-colMeans(sand.obs))^2)/colSums((sand.obs-
colMeans(sand.obs))^2)
sand.r2aj=as.numeric((r2aj(sand.r2,150,4))[[1]])
sand.int=(colMaxs(sand.obs)-colMins(sand.obs))
sand.rmse.n=sand.rmse/sand.int
sand.stats=cbind(sand.rmse,sand.rmse.n,sand.me,sand.r2aj)
mediana=matrix(c(apply(sand.stats,2,median)),nrow=4)
percentis=t(matrix(c(quantile(sand.rmse,c(0.025,0.975)),quantile(sand.rmse.
n,c(0.025,0.975)),quantile(sand.me,c(0.025,0.975)),
quantile(sand.r2aj,c(0.025,0.975))),nrow=2))
sand.stats=cbind(percentis[,1],mediana,percentis[,2]);colnames(sand.stats)=
c("2,5 percentil","Mediana","97,5
percentil");rownames(sand.stats)=c("RMSE","RMSEnorm","EM","R2aj")

# Cálculo das estatísticas da predição do silte
silt.res=silt.pred-silt.obs
silt.me=colMeans(silt.res)
silt.eqm=colMeans(silt.res^2)
silt.rmse=sqrt(silt.eqm)
silt.r2=colSums((silt.pred-colMeans(silt.obs))^2)/colSums((silt.obs-
colMeans(silt.obs))^2)
silt.r2aj=as.numeric((r2aj(silt.r2,150,4))[[1]])
silt.int=(colMaxs(silt.obs)-colMins(silt.obs))
silt.rmse.n=silt.rmse/silt.int
silt.stats=cbind(silt.rmse,silt.rmse.n,silt.me,silt.r2aj)
mediana=matrix(c(apply(silt.stats,2,median)),nrow=4)
percentis=t(matrix(c(quantile(silt.rmse,c(0.025,0.975)),quantile(silt.rmse.
n,c(0.025,0.975)),quantile(silt.me,c(0.025,0.975)),
quantile(silt.r2aj,c(0.025,0.975))),nrow=2))
```

```
silt.stats=cbind(percentis[,1],mediana,percentis[,2]);colnames(silt.stats)=
c("2,5 percentil","Mediana","97,5
percentil");rownames(silt.stats)=c("RMSE","RMSEnorm","EM","R2aj")

# Cálculo das estatísticas da predição da argila
clay.res=clay.pred-clay.obs
clay.me=colMeans(clay.res)
clay.eqm=colMeans(clay.res^2)
clay.rmse=sqrt(clay.eqm)
clay.r2=colSums((clay.pred-colMeans(clay.obs))^2)/colSums((clay.obs-
colMeans(clay.obs))^2)
clay.r2aj=as.numeric((r2aj(clay.r2,150,4))[[1]])
clay.int=(colMaxs(clay.obs)-colMins(clay.obs))
clay.rmse.n=clay.rmse/clay.int
clay.stats=cbind(clay.rmse,clay.rmse.n,clay.me,clay.r2aj)
mediana=matrix(c(apply(clay.stats,2,median)),nrow=4)
percentis=t(matrix(c(quantile(clay.rmse,c(0.025,0.975)),quantile(clay.rmse.
n,c(0.025,0.975)),quantile(clay.me,c(0.025,0.975)),
quantile(clay.r2aj,c(0.025,0.975))),nrow=2))
clay.stats=cbind(percentis[,1],mediana,percentis[,2]);colnames(clay.stats)=
c("2,5 percentil","Mediana","97,5
percentil");rownames(clay.stats)=c("RMSE","RMSEnorm","EM","R2aj")

# Estatísticas da validação cruzada
sand.stats
silt.stats
clay.stats

# Validação cruzada no domínio fisiográfico superior
attach(dados_sup)
preditoras1=as.data.frame(cbind(ELEV,IPE))
ln.clay.sand=log(clay/sand)
preditoras2=as.data.frame(cbind(ELEV))
ln.silt.sand=log(silt/sand)
detach(dados_sup)

# validação cruzada da FPESe para estimar ln(clay/sand)
set.seed(123)
tmp.clay.sand=createMultiFolds(ln.clay.sand,k=10,times=100)
trControl.clay.sand=trainControl(method="repeatedcv",number=10,repeats=100,
savePredictions=TRUE,index=tmp.clay.sand)
cv.clay.sand=train(preditoras1,ln.clay.sand,"lm",trControl=trControl.clay.s
and)

# identificação das observações usadas como conjunto de teste
set.seed(123)
cv.clay.sand.ID=obs.id(ln.clay.sand,k=10,times=100)
L.max <- max(sapply(cv.clay.sand.ID, length), na.rm = TRUE)
cv.clay.sand.ID =t(sapply(cv.clay.sand.ID, function(x) c(x, rep(NA, L.max -
length(x)))))
cv.clay.sand.ID =as.numeric(na.omit(c(t(as.data.frame(cv.clay.sand.ID)))))
```

```
# organização dos dados para cálculo das estatísticas da validação cruzada
cv.clay.sand=as.data.frame(cv.clay.sand$pred[1:2])
cv.rep=sort(rep(1:100,174))
cv.clay.sand=cbind(cv.rep,cv.clay.sand.ID,cv.clay.sand)
rm(cv.clay.sand.ID, cv.rep)
attach(cv.clay.sand)
cv.clay.sand=cv.clay.sand[order(cv.rep,cv.clay.sand.ID),]
detach(cv.clay.sand)

# validação cruzada da FPESe para estimar ln(silt/sand)
set.seed(123)
tmp.silt.sand=createMultiFolds(ln.silt.sand,k=10,times=100)
trControl.silt.sand=trainControl(method="repeatedcv",number=10,repeats=100,
savePredictions=TRUE,index=tmp.silt.sand)
cv.silt.sand=train(preditoras2,ln.silt.sand,"lm",trControl=trControl.silt.s
and)

# identificação das observações usadas como conjunto de teste
set.seed(123)
cv.silt.sand.ID=obs.id(ln.silt.sand,k=10,times=100)
L.max <- max(sapply(cv.silt.sand.ID, length), na.rm = TRUE)
cv.silt.sand.ID =t(sapply(cv.silt.sand.ID, function(x) c(x, rep(NA, L.max -
length(x)))))
cv.silt.sand.ID =as.numeric(na.omit(c(t(as.data.frame(cv.silt.sand.ID)))))

# organização dos dados para cálculo das estatísticas da validação cruzada
cv.silt.sand=as.data.frame(cv.silt.sand$pred[1:2])
cv.rep=sort(rep(1:100,174))
cv.silt.sand=cbind(cv.rep,cv.silt.sand.ID,cv.silt.sand)
rm(cv.silt.sand.ID, cv.rep)
attach(cv.silt.sand)
cv.silt.sand=cv.silt.sand[order(cv.rep,cv.silt.sand.ID),]
detach(cv.silt.sand)

# cálculo das estatísticas da validação cruzada
sand.obs=(         1
/(exp(cv.clay.sand$obs)+exp(cv.silt.sand$obs)+1))*100
silt.obs=((exp(cv.silt.sand$obs))/(exp(cv.clay.sand$obs)+exp(cv.silt.sand$o
bs)+1))*100
clay.obs=((exp(cv.clay.sand$obs))/(exp(cv.clay.sand$obs)+exp(cv.silt.sand$o
bs)+1))*100
sand.obs=matrix(sand.obs,nrow=174)
silt.obs=matrix(silt.obs,nrow=174)
clay.obs=matrix(clay.obs,nrow=174)
sand.pred=(1/(exp(cv.clay.sand$pred)+exp(cv.silt.sand$pred)+1))*100
silt.pred=(exp(cv.silt.sand$pred)/(exp(cv.clay.sand$pred)+exp(cv.silt.sand$
pred)+1))*100
clay.pred=(exp(cv.clay.sand$pred)/(exp(cv.clay.sand$pred)+exp(cv.silt.sand$
pred)+1))*100
sand.pred=matrix(sand.pred,nrow=174)
silt.pred=matrix(silt.pred,nrow=174)
clay.pred=matrix(clay.pred,nrow=174)
```

```
sand.res=sand.pred-sand.obs
sand.me=colMeans(sand.res)
sand.eqm=colMeans(sand.res^2)
sand.rmse=sqrt(sand.eqm)
sand.r2=colSums((sand.pred-colMeans(sand.obs))^2)/colSums((sand.obs-
colMeans(sand.obs))^2)
sand.r2aj=as.numeric((r2aj(sand.r2,150,3))[[1]])
sand.int=(colMaxs(sand.obs)-colMins(sand.obs))
sand.rmse.n=sand.rmse/sand.int
sand.stats=cbind(sand.rmse,sand.rmse.n,sand.me,sand.r2aj)
mediana=matrix(c(apply(sand.stats,2,median)),nrow=4)
percentis=t(matrix(c(quantile(sand.rmse,c(0.025,0.975)),quantile(sand.rmse.
n,c(0.025,0.975)),quantile(sand.me,c(0.025,0.975)),
quantile(sand.r2aj,c(0.025,0.975))),nrow=2))
sand.stats=cbind(percentis[,1],mediana,percentis[,2]);colnames(sand.stats)=
c("2,5 percentil","Mediana","97,5
percentil");rownames(sand.stats)=c("RMSE","RMSEnorm","EM","R2aj")

silt.res=silt.pred-silt.obs
silt.me=colMeans(silt.res)
silt.eqm=colMeans(silt.res^2)
silt.rmse=sqrt(silt.eqm)
silt.r2=colSums((silt.pred-colMeans(silt.obs))^2)/colSums((silt.obs-
colMeans(silt.obs))^2)
silt.r2aj=as.numeric((r2aj(silt.r2,150,3))[[1]])
silt.int=(colMaxs(silt.obs)-colMins(silt.obs))
silt.rmse.n=silt.rmse/silt.int
silt.stats=cbind(silt.rmse,silt.rmse.n,silt.me,silt.r2aj)
mediana=matrix(c(apply(silt.stats,2,median)),nrow=4)
percentis=t(matrix(c(quantile(silt.rmse,c(0.025,0.975)),quantile(silt.rmse.
n,c(0.025,0.975)),quantile(silt.me,c(0.025,0.975)),
quantile(silt.r2aj,c(0.025,0.975))),nrow=2))
silt.stats=cbind(percentis[,1],mediana,percentis[,2]);colnames(silt.stats)=
c("2,5 percentil","Mediana","97,5
percentil");rownames(silt.stats)=c("RMSE","RMSEnorm","EM","R2aj")

clay.res=clay.pred-clay.obs
clay.me=colMeans(clay.res)
clay.eqm=colMeans(clay.res^2)
clay.rmse=sqrt(clay.eqm)
clay.r2=colSums((clay.pred-colMeans(clay.obs))^2)/colSums((clay.obs-
colMeans(clay.obs))^2)
clay.r2aj=as.numeric((r2aj(clay.r2,150,3))[[1]])
clay.int=(colMaxs(clay.obs)-colMins(clay.obs))
clay.rmse.n=clay.rmse/clay.int
clay.stats=cbind(clay.rmse,clay.rmse.n,clay.me,clay.r2aj)
mediana=matrix(c(apply(clay.stats,2,median)),nrow=4)
percentis=t(matrix(c(quantile(clay.rmse,c(0.025,0.975)),quantile(clay.rmse.
n,c(0.025,0.975)),quantile(clay.me,c(0.025,0.975)),
quantile(clay.r2aj,c(0.025,0.975))),nrow=2))
clay.stats=cbind(percentis[,1],mediana,percentis[,2]);colnames(clay.stats)=
c("2,5 percentil","Mediana","97,5
percentil");rownames(clay.stats)=c("RMSE","RMSEnorm","EM","R2aj")
```

```
# estatísticas da validação cruzada
sand.stats
silt.stats
clay.stats

# validação cruzada das FPESe construídas para toda a área de estudo em
cada um dos domínios fisiográficos
# definição os conjuntos de dados de variáveis dependentes e preditoras
attach(dados)
preditoras=cbind(CONV,ELEV,IPE)
ln.clay.sand=log(clay/sand)
ln.silt.sand=log(silt/sand)
detach(dados)

# validação cruzada da FPESe para estimar ln(clay/sand) em toda a área de
estudo
set.seed(123)
tmp.clay.sand=createMultiFolds(ln.clay.sand,k=10,times=100)
trControl.clay.sand=trainControl(method="repeatedcv",number=10,repeats=100,
savePredictions=TRUE,index=tmp.clay.sand)
cv.clay.sand=train(preditoras,ln.clay.sand,"lm",trControl=trControl.clay.sa
nd)

# identificação das observações usadas como conjunto de teste
set.seed(123)
cv.clay.sand.ID=obs.id(ln.clay.sand,k=10,times=100)
L.max <- max(sapply(cv.clay.sand.ID, length), na.rm = TRUE)
cv.clay.sand.ID =t(sapply(cv.clay.sand.ID, function(x) c(x, rep(NA, L.max -
length(x)))))
cv.clay.sand.ID =as.numeric(na.omit(c(t(as.data.frame(cv.clay.sand.ID)))))

# Organização dos dados
cv.clay.sand=as.data.frame(cv.clay.sand$pred[1:2])
cv.rep=sort(rep(1:100,339))
cv.clay.sand=cbind(cv.rep,cv.clay.sand.ID,cv.clay.sand)
rm(cv.clay.sand.ID, cv.rep)
attach(cv.clay.sand)
cv.clay.sand=cv.clay.sand[order(cv.rep,cv.clay.sand.ID),]
detach(cv.clay.sand)

# validação cruzada da FPESe para estimar ln(silt/sand) em toda a área de
estudo
set.seed(123)
tmp.silt.sand=createMultiFolds(ln.silt.sand,k=10,times=100)
trControl.silt.sand=trainControl(method="repeatedcv",number=10,repeats=100,
savePredictions=TRUE,index=tmp.silt.sand)
cv.silt.sand=train(preditoras,ln.silt.sand,"lm",trControl=trControl.silt.sa
nd)

# identificação das observações usadas como conjunto de teste
set.seed(123)
cv.silt.sand.ID=obs.id(ln.silt.sand,k=10,times=100)
L.max <- max(sapply(cv.silt.sand.ID, length), na.rm = TRUE)
```

```
cv.silt.sand.ID =t(sapply(cv.silt.sand.ID, function(x) c(x, rep(NA, L.max -
length(x)))))
cv.silt.sand.ID =as.numeric(na.omit(c(t(as.data.frame(cv.silt.sand.ID)))))

# Organização dos dados
cv.silt.sand=as.data.frame(cv.silt.sand$pred[1:2])
cv.rep=sort(rep(1:100,339))
cv.silt.sand=cbind(cv.rep,cv.silt.sand.ID,cv.silt.sand)
rm(cv.silt.sand.ID, cv.rep)
attach(cv.silt.sand)
cv.silt.sand=cv.silt.sand[order(cv.rep,cv.silt.sand.ID),]
detach(cv.silt.sand)
sand.obs=(         1
/(exp(cv.clay.sand$obs)+exp(cv.silt.sand$obs)+1))*100
silt.obs=((exp(cv.silt.sand$obs))/(exp(cv.clay.sand$obs)+exp(cv.silt.sand$o
bs)+1))*100
clay.obs=((exp(cv.clay.sand$obs))/(exp(cv.clay.sand$obs)+exp(cv.silt.sand$o
bs)+1))*100
sand.obs=matrix(sand.obs,nrow=339)
silt.obs=matrix(silt.obs,nrow=339)
clay.obs=matrix(clay.obs,nrow=339)
sand.pred=(1/(exp(cv.clay.sand$pred)+exp(cv.silt.sand$pred)+1))*100
silt.pred=(exp(cv.silt.sand$pred)/(exp(cv.clay.sand$pred)+exp(cv.silt.sand$
pred)+1))*100
clay.pred=(exp(cv.clay.sand$pred)/(exp(cv.clay.sand$pred)+exp(cv.silt.sand$
pred)+1))*100
sand.pred=matrix(sand.pred,nrow=339)
silt.pred=matrix(silt.pred,nrow=339)
clay.pred=matrix(clay.pred,nrow=339)

attach(dados)
sand.obs=cbind(sand.obs,ELEV)
silt.obs=cbind(silt.obs,ELEV)
clay.obs=cbind(clay.obs,ELEV)
sand.pred=cbind(sand.pred,ELEV)
silt.pred=cbind(silt.pred,ELEV)
clay.pred=cbind(clay.pred,ELEV)
detach(dados)
sand.obs= as.data.frame(sand.obs)
silt.obs= as.data.frame(silt.obs)
clay.obs= as.data.frame(clay.obs)
sand.pred= as.data.frame(sand.pred)
silt.pred= as.data.frame(silt.pred)
clay.pred= as.data.frame(clay.pred)

# domínio inferior
sand.pred.inf=sand.pred[which(sand.pred$ELEV<300),]
sand.pred.inf$ELEV=NULL
silt.pred.inf=silt.pred[which(silt.pred$ELEV<300),]
silt.pred.inf$ELEV=NULL
clay.pred.inf=clay.pred[which(clay.pred$ELEV<300),]
clay.pred.inf$ELEV=NULL
sand.obs.inf=sand.obs[which(sand.obs$ELEV<300),]
```

```
sand.obs.inf$ELEV=NULL
silt.obs.inf=silt.obs[which(silt.obs$ELEV<300),]
silt.obs.inf$ELEV=NULL
clay.obs.inf=clay.obs[which(clay.obs$ELEV<300),]
clay.obs.inf$ELEV=NULL

sand.res=sand.pred.inf-sand.obs.inf
sand.me=colMeans(sand.res)
sand.eqm=colMeans(sand.res^2)
sand.rmse=sqrt(sand.eqm)
sand.r2=colSums((sand.pred.inf-
colMeans(sand.obs.inf))^2)/colSums((sand.obs.inf-colMeans(sand.obs.inf))^2)
sand.r2aj=as.numeric((r2aj(sand.r2,300,4))[[1]])
sand.int=(colMaxs(sand.obs.inf)-colMins(sand.obs.inf))
sand.rmse.n=sand.rmse/sand.int
sand.stats=cbind(sand.rmse,sand.rmse.n,sand.me,sand.r2aj)
mediana=matrix(c(apply(sand.stats,2,median)),nrow=4)
percentis=t(matrix(c(quantile(sand.rmse,c(0.025,0.975)),quantile(sand.rmse.
n,c(0.025,0.975)),quantile(sand.me,c(0.025,0.975)),
quantile(sand.r2aj,c(0.025,0.975))),nrow=2))
sand.stats=cbind(percentis[,1],mediana,percentis[,2]);colnames(sand.stats)=
c("2,5 percentil","Mediana","97,5
percentil");rownames(sand.stats)=c("RMSE","RMSEnorm","EM","R2aj")

silt.res=silt.pred.inf-silt.obs.inf
silt.me=colMeans(silt.res)
silt.eqm=colMeans(silt.res^2)
silt.rmse=sqrt(silt.eqm)
silt.r2=colSums((silt.pred.inf-
colMeans(silt.obs.inf))^2)/colSums((silt.obs.inf-colMeans(silt.obs.inf))^2)
silt.r2aj=as.numeric((r2aj(silt.r2,300,4))[[1]])
silt.int=(colMaxs(silt.obs.inf)-colMins(silt.obs.inf))
silt.rmse.n=silt.rmse/silt.int
silt.stats=cbind(silt.rmse,silt.rmse.n,silt.me,silt.r2aj)
mediana=matrix(c(apply(silt.stats,2,median)),nrow=4)
percentis=t(matrix(c(quantile(silt.rmse,c(0.025,0.975)),quantile(silt.rmse.
n,c(0.025,0.975)),quantile(silt.me,c(0.025,0.975)),
quantile(silt.r2aj,c(0.025,0.975))),nrow=2))
silt.stats=cbind(percentis[,1],mediana,percentis[,2]);colnames(silt.stats)=
c("2,5 percentil","Mediana","97,5
percentil");rownames(silt.stats)=c("RMSE","RMSEnorm","EM","R2aj")

clay.res=clay.pred.inf-clay.obs.inf
clay.me=colMeans(clay.res)
clay.eqm=colMeans(clay.res^2)
clay.rmse=sqrt(clay.eqm)
clay.r2=colSums((clay.pred.inf-
colMeans(clay.obs.inf))^2)/colSums((clay.obs.inf-colMeans(clay.obs.inf))^2)
clay.r2aj=as.numeric((r2aj(clay.r2,300,4))[[1]])
clay.int=(colMaxs(clay.obs.inf)-colMins(clay.obs.inf))
clay.rmse.n=clay.rmse/clay.int
clay.stats=cbind(clay.rmse,clay.rmse.n,clay.me,clay.r2aj)
mediana=matrix(c(apply(clay.stats,2,median)),nrow=4)
```

```
percentis=t(matrix(c(quantile(clay.rmse,c(0.025,0.975)),quantile(clay.rmse.
n,c(0.025,0.975)),quantile(clay.me,c(0.025,0.975)),
quantile(clay.r2aj,c(0.025,0.975))),nrow=2))
clay.stats=cbind(percentis[,1],mediana,percentis[,2]);colnames(clay.stats)=
c("2,5 percentil","Mediana","97,5
percentil");rownames(clay.stats)=c("RMSE","RMSEnorm","EM","R2aj")

# estatísticas da validação cruzada no domínio inferior
sand.stats
silt.stats
clay.stats

# domínio superior
sand.pred.sup=sand.pred[which(sand.pred$ELEV>=300),]
sand.pred.sup$ELEV=NULL
silt.pred.sup=silt.pred[which(silt.pred$ELEV>=300),]
silt.pred.sup$ELEV=NULL
clay.pred.sup=clay.pred[which(clay.pred$ELEV>=300),]
clay.pred.sup$ELEV=NULL
sand.obs.sup=sand.obs[which(sand.obs$ELEV>=300),]
sand.obs.sup$ELEV=NULL
silt.obs.sup=silt.obs[which(silt.obs$ELEV>=300),]
silt.obs.sup$ELEV=NULL
clay.obs.sup=clay.obs[which(clay.obs$ELEV>=300),]
clay.obs.sup$ELEV=NULL

sand.res=sand.pred.sup-sand.obs.sup
sand.me=colMeans(sand.res)
sand.eqm=colMeans(sand.res^2)
sand.rmse=sqrt(sand.eqm)
sand.r2=colSums((sand.pred.sup-
colMeans(sand.obs.sup))^2)/colSums((sand.obs.sup-colMeans(sand.obs.sup))^2)
sand.r2aj=as.numeric((r2aj(sand.r2,300,4))[[1]])
sand.int=(colMaxs(sand.obs.sup)-colMins(sand.obs.sup))
sand.rmse.n=sand.rmse/sand.int
sand.stats=cbind(sand.rmse,sand.rmse.n,sand.me,sand.r2aj)
mediana=matrix(c(apply(sand.stats,2,median)),nrow=4)
percentis=t(matrix(c(quantile(sand.rmse,c(0.025,0.975)),quantile(sand.rmse.
n,c(0.025,0.975)),quantile(sand.me,c(0.025,0.975)),
quantile(sand.r2aj,c(0.025,0.975))),nrow=2))
sand.stats=cbind(percentis[,1],mediana,percentis[,2]);colnames(sand.stats)=
c("2,5 percentil","Mediana","97,5
percentil");rownames(sand.stats)=c("RMSE","RMSEnorm","EM","R2aj")

silt.res=silt.pred.sup-silt.obs.sup
silt.me=colMeans(silt.res)
silt.eqm=colMeans(silt.res^2)
silt.rmse=sqrt(silt.eqm)
silt.r2=colSums((silt.pred.sup-
colMeans(silt.obs.sup))^2)/colSums((silt.obs.sup-colMeans(silt.obs.sup))^2)
silt.r2aj=as.numeric((r2aj(silt.r2,300,4))[[1]])
silt.int=(colMaxs(silt.obs.sup)-colMins(silt.obs.sup))
silt.rmse.n=silt.rmse/silt.int
```

```
silt.stats=cbind(silt.rmse,silt.rmse.n,silt.me,silt.r2aj)
mediana=matrix(c(apply(silt.stats,2,median)),nrow=4)
percentis=t(matrix(c(quantile(silt.rmse,c(0.025,0.975)),quantile(silt.rmse.
n,c(0.025,0.975)),quantile(silt.me,c(0.025,0.975)),
quantile(silt.r2aj,c(0.025,0.975))),nrow=2))
silt.stats=cbind(percentis[,1],mediana,percentis[,2]);colnames(silt.stats)=
c("2,5 percentil","Mediana","97,5
percentil");rownames(silt.stats)=c("RMSE","RMSEnorm","EM","R2aj")

clay.res=clay.pred.sup-clay.obs.sup
clay.me=colMeans(clay.res)
clay.eqm=colMeans(clay.res^2)
clay.rmse=sqrt(clay.eqm)
clay.r2=colSums((clay.pred.sup-
colMeans(clay.obs.sup))^2)/colSums((clay.obs.sup-colMeans(clay.obs.sup))^2)
clay.r2aj=as.numeric((r2aj(clay.r2,300,4))[[1]])
clay.int=(colMaxs(clay.obs.sup)-colMins(clay.obs.sup))
clay.rmse.n=clay.rmse/clay.int
clay.stats=cbind(clay.rmse,clay.rmse.n,clay.me,clay.r2aj)
mediana=matrix(c(apply(clay.stats,2,median)),nrow=4)
percentis=t(matrix(c(quantile(clay.rmse,c(0.025,0.975)),quantile(clay.rmse.
n,c(0.025,0.975)),quantile(clay.me,c(0.025,0.975)),
quantile(clay.r2aj,c(0.025,0.975))),nrow=2))
clay.stats=cbind(percentis[,1],mediana,percentis[,2]);colnames(clay.stats)=
c("2,5 percentil","Mediana","97,5
percentil");rownames(clay.stats)=c("RMSE","RMSEnorm","EM","R2aj")

# estatísticas da validação cruzada no domínio inferior
sand.stats
silt.stats
clay.stats
```