BRUNO NERY FERNANDES VASCONCELOS

MAPEAMENTO DIGITAL DE SOLOS EM DIFERENTES ESCALAS: ABORDAGEM METODOLÓGICA

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Solos e Nutrição de Plantas, para obtenção do título de Doctor Scientiae.

VIÇOSA
MINAS GERAIS – BRASIL
2016

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
V331m
2016

Vasconcelos, Bruno Nery Fernandes, 1981-
Mapeamento digital de solos em diferentes escalas : abordagem metodológica / Bruno Nery Fernandes Vasconcelos. – Viçosa, MG, 2016.
x, 107f. : il. (algumas color.) ; 29 cm.

Orientador: Elpídio Inácio Fernandes Filho.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Mapeamento de solos. 2. Levantamentos de solos. 3. Solos - Mapeamento digital. I. Universidade Federal de Viçosa. Departamento de Solos. Programa de Pós-graduação em Solos e Nutrição de Plantas. II. Título.

CDD 22. ed. 631.47

BRUNO NERY FERNANDES VASCONCELOS

MAPEAMENTO DIGITAL DE SOLOS EM DIFERENTES ESCALAS: ABORDAGEM METODOLÓGICA

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Solos e Nutrição de Plantas, para obtenção do título de *Doctor Scientiae.*

APROVADA: 13 de setembro de 2016.

Cesar da Silva Chagas

Bruno Araújo Furtado de Mendonça

Márcio Rocha Francelino

Carlos Ernesto G.R. Schaefer
(Coorientador)

Elpídio Inácio Fernandes Filho
(Orientador)

AGRADECIMENTOS

Ao **Pai Celestial** pela graça da Vida junto com todas as oportunidades que vem junto com ela;

A **Mãe Divina** pela companhia inseparável por toda a caminhada;

Ao Nosso Senhor **Jesus Cristo** pelo seu Amor Incondicional: Caminho, Verdade e Vida;

Aos meus pais **Wilton** e **Ludmila**, pelo amor, paciência e dedicação na instrução que me deram nesta passagem pela Terra;

A minha irmã **Marina** pelo exemplo e por sua existência em minha vida;

Aos meus filhos **Antônio** e **Maria** por me possibilitarem a experiência mais profunda e pura de Amor que pude vivenciar na minha existência;

A Viçosa, pelos maravilhosos anos de vivência e pelas incontáveis oportunidades de crescimento e aprendizado, juntamente com as pessoas maravilhosas que conheci;

Aos professores do Departamento de Solos da Universidade Federal de Viçosa, em especial: ao professor João Carlos Ker, que acreditou em meu potencial desde o início da minha caminhada acadêmica, estendo sua mão e me instruindo nos primeiros passos desta caminhada; ao professor Carlos Ernesto G.R. Schaefer pelas oportunidades e compartilhamento de seu extenso conhecimento; ao professor Elpídio Inácio Fernandes Filho, um exemplo de orientador, sempre disposto a discutir idéias e compartilhar inúmeras horas de seu tempo para atender as pessoas que o procuram;

Aos colegas de pós gradução do Departamento de Solos, em especial aos amigos do Laboratório de Geoprocessamento (LabGeo);

Aos funcionários do departmento de solos e da Universidade que executando suas tarefas diárias auxiliam no funcionamento desta enorme engrenagem. Em especial às secretárias Luciana e Rita que sempre me auxiliaram com boa vontade a tramitar pelos caminhos, muitas vezes tortuosos, desta mesma engrenagem;

Aos membros da banca de avaliação deste trabalho que contribuíram com a melhoria do mesmo, a partir de suas considerações;

Ao povo brasileiro que por meio do CNPq e da CAPES proporcionaram o financiamento destes meus anos de estudo;

A todos que buscam a cada dia viver melhor aproveitando seu tempo para aprender a amar mais, neste maravilhoso planeta escola.

GRATIDÃO!!!

BIOGRAFIA

BRUNO NERY FERNANDES VASCONCELOS, filho de Wilton Fernandes Vasconcelos e Ludmila Aparecida Nery Vasconcelos, pai de Maria Nery Santos Vasconcelos e de Antônio Campos Vasconcelos, nasceu em 20 de fevereiro de 1981, na cidade de Belo Horizonte, Minas Gerais.

Em 2001 iniciou o Curso de Agronomia na Universidade Federal de Viçosa, Viçosa, Minas Gerais. Ao longo da graduação desenvolveu algumas atividades no departamento de solos como monitoria, iniciação científica e participação em trabalhos de levantamento de solos. Também realizou diversas atividades extra-acadêmicas tais como, participação em grupos de agroecologia, participação em eventos (congressos, simpósios, minicursos etc...), e na elaboração de atividades práticas principalmente voltadas para a agricultura, desenvolvendo assim uma capacidade melhor de trabalhar em grupo e uma percepção multidisciplinar.

Em março de 2008 iniciou o Curso de Mestrado no Programa de Pós-Graduação em Solos e Nutrição de Plantas da Universidade Federal de Viçosa finalizando-o com a defesa da presente dissertação, em 22 de março de 2010.

Após alguns trabalhos práticos de levantamentos de solos em Viçosa, no período de 2010 a 2013 vivenciou outras realidades em algumas regiões de Minas Gerais tais como: Alto Paranaíba, Quadrilátero Ferrífero e Sul de Minas, retornando em meados de 2013 para cursar o Doutorado no Programa de Pós-Graduação em Solos e Nutrição de Plantas.

Após a conclusão do Doutorado prepara-se para ingressar na carreira de docente na Universidade Federal de Uberlândia, Instituto de Ciências Agrárias-Campus Monte Carmelo.

SUMÁRIO

CAPÍTULO 2 - AVALIAÇÃO DE DADOS LEGADOS NO BALANCEAMENTO DE AMOSTRAGEM COM HIPERCUBO LATINO CONDICIONADO PARA MAPEAMENTO DIGITAL DE CLASSES DE SOLO

CAPITULO 3 - SELEÇÃO DE COVARIÁVEIS PARA PREDIÇÃO DE UNIDADES DE MAPEAMENTO DE SOLOS NO QUADRILÁTERO FERRÍFERO, MINAS GERAIS

RESUMO

VASCONCELOS, Bruno Nery Fernandes, D.Sc., Universidade Federal de Viçosa, setembro de 2016.**Mapeamento digital de solos em diferentes escalas: abordagem metodológica.** Orientador: Elpídio Inácio Fernandes Filho. Coorientadores: João Carlos ker e Carlos Ernesto G.R. Schaefer.

A demanda por informações mais detalhadas a cerca do recurso de solos vem crescendo vertiginosamente, frente a uma necessidade de planejamento mais eficiente de uso deste recurso. O levantamento de solos permite a descrição das características físicas, químicas e mineralógicas dos solos de uma área, delimitando suas ocorrências na paisagem, através de mapas. Novos desafios estão sendo relacionados aos levantamentos de solos, como resultados do rápido desenvolvimento das técnicas de Sistema de Informação Geográfica e Sensoriamento Remoto, exigindo que os pedólogos adotem novas técnicas para tornar os levantamentos mais rápidos, menos onerosos e mais quantitativos. Neste contexto surge o Mapeamento Digital de Solos (MDS) que pode ser definido como a criação de um sistema espacial de distribuição dos solos por modelos numéricos, que permitem inferir sobre as variações espaciais e temporais dos tipos de solos e de suas propriedades. No entanto esta técnica se encontra em fase de consolidação e aprimoramento e metodológica, apresentando diversas lacunas de conhecimento que precisam ser pesquisadas. O presente trabalho tem como objetivo central propor e avaliar metodologias focadas nas questões de amostragem e de seleção de covariáveis ambientais. Para tanto, optou-se por trabalhar em três áreas com distintas proporções, contemplando assim cenários de mapeamentos locais, regionais e nacionais. Foi elaborada uma proposta de balanceamento de amostragem, através do uso de dados legados em uma modificação do método de amostragem Hipercubo Latino Condicionado (cLHS). Esta proposta foi testada no Estado de Minas Gerais e na Bacia do Rio Turvo Sujo-MG. A seleção de covariáveis ambientais foi avaliada, a partir de quatro formas diferentes, na região do Quadrilátero Ferrífero-MG. Os resultados evidenciam que tanto o balanceamento amostral, quanto a seleção de covariáveis são

processos fundamentais no Mapeamento Digital de Solos, que devem ser adotadas como etapas intrínsecas neste procedimento metodológico.

ABSTRACT

VASCONCELOS, Bruno Nery Fernandes, D.Sc., Universidade Federal de Viçosa, September, 2016. **Digital soil mapping at different scales: methodological approach.** Adviser: Elpídio Inácio Fernandes Filho. Co-advisers: João Carlos Ker and Carlos Ernesto G.R. Schaefer.

The demand for more detailed information about the soil resource is growing dramatically, due to the need of more efficient planning use of this feature. Soil survey allows description of the physical, chemical and mineralogical soil characteristics of an area, limiting their occurrence in a landscape, through maps. New challenges are related to soil surveys, as a result of the rapid development of Geographic Information System and Remote Sensing techniques, requiring from soil scientists to adopt new techniques to make the surveys faster, less expensive and more quantitative. In this context the Digital Soil Mapping (DSM) arises which can be defined as the creation of a spatial system through numerical models for soils distribution, which allows inferring the spatial and temporal variations of soil types and their properties. However this technique is yet in consolidation and improvement of the methodology, with many knowledge gaps that need to be researched. This study aimed to propose and evaluate methodologies focused on sampling issues and selection of environmental covariates. Therefore, it was choosen to work in three areas with different proportions, contemplating local, regional and national scenarios mappings. A sample balancing proposal was developed through the use of legacy data in a modification of the Contidioned Latin Hypercube (cLHS) sampling method. This proposal was tested for the Minas Gerais State and for the Turvo Sujo River Basin. The selection of environmental covariates was evaluated through four different ways in the Quadrilátero Ferrífero (Iron Quadrangle) Region. The results show that both the sample balancing, and the selection of covariates are critical processes in the Digital Soil Mapping, which should be adopted as intrinsic methodological steps in this procedure.

1. INTRODUÇÃO GERAL

O conhecimento pormenorizado dos solos é essencial para qualquer civilização que almeje um crescimento socioeconômico equilibrado. Os levantamentos de solos são a metodologia mais comumente utilizada para obter tais informações, e constituem a base fundamental para a composição de unidades de mapeamento, cuja distribuição geográfica, extensão e limites são evidenciados em mapas (EMBRAPA, 1995; IBGE, 2015).

Um levantamento de solos através da avaliação das características morfológicas, físicas, químicas e mineralógicas dos solos de uma área, permite o enquadramento dos mesmos segundo um sistema de classificação, além de delimitar sua ocorrência na paisagem, com possibilidades de posterior interpretação para uso agrícola e não agrícola (USDA, 1993; EMBRAPA, 1995; IBGE, 2015; Resende et al., 2007).

No Brasil, desde a década de cinquenta, quando surgiram os primeiros levantamentos de solos, a maioria desses mapeamentos foi realizada com escalas pouco detalhadas, em nível exploratório ou de reconhecimento de baixa intensidade. Desta forma extensas áreas em todo o país ainda carecem de levantamentos mais detalhados para subsidiar decisões de uso e conservação dos solos (MENDONÇA-SANTOS e SANTOS, 2007).

A década de 70 pode ser considerada como o período áureo dos levantamentos de solos no Brasil, onde mais se produziu informação e resultados a respeito deste tema. Foi nesta fase que foi criado o Serviço Nacional de Levantamento e Conservação de Solos (SNLCS), que conduziu trabalhos importantes como o mapeamento de caráter exploratório ou reconhecimento de baixa intensidade de todos os Estados do Nordeste e da região norte do Estado de Minas Gerais. Foi nesta época também, que se iniciaram as atividades do Projeto Radam (Radar da Amazônia), posteriormente denominado de RadamBrasil, que mapeou praticamente todo o território nacional na escala de 1:1.000.000, até o início da década de 80, onde o mesmo foi concluído. O acervo de informações geradas permitiu que o SNLCS estruturasse o mapa de solos do Brasil na escala de 1:5.000.000 que foi publicado em 1981, e permitiu uma visão panorâmica da diversidade dos recursos de solos no país.

No início dos anos 90, após ser extinto o SNLCS, iniciou-se uma fase de declínio na produção de levantamentos de solos no Brasil. No ano de 2013 o grupo de Geomática da Embrapa Solos (CNPS) adicionou outras informações às já existentes e conclui que os levantamentos de caráter exploratório contemplavam 94% do território nacional, seguidos pelos levantamentos de reconhecimento de baixa intensidade que alcançavam 84,2% do mesmo. Por outro lado, os levantamentos semidetalhados (escalas entre 1:25.000 e 1:50.000) e detalhados (escalas entre 1:7.000 a 1:15.000) ocupam respectivamente apenas 0,61% e 0,00003% da área total do país (OLIVEIRA, 2014).

Conforme o panorama dos mapeamentos de solo no Brasil pode-se verificar que extensas áreas em todo o país ainda carecem de mapeamentos com maior nível de detalhamento que permitam planejamentos em escala regional, adequada para solucionar problemas de uso do solo, manejo, conservação, prevenção e recuperação de áreas degradadas, agrícolas e não agrícolas (CHAGAS, 2006; KER, 2007; OLIVEIRA, 2007; SANTOS, 2007). Desta forma, a atualização dos mapeamentos de solos, em maiores níveis de detalhamento, é uma demanda urgente para o planejamento agrícola e ambiental, o que torna este tipo de atividade uma necessidade contínua (BASHER, 1997).

Atualmente alguns desafios estão sendo relacionados aos levantamentos de solos, como resultados do rápido desenvolvimento dos Sistemas de Informação Geográfica (SIG's), apontando para uma necessidade de adequação das técnicas convencionais de mapeamento em relação a essas novas tendências. Chagas (2006) ressalta que os pedólogos devem buscar, por meio de pesquisa, a adoção de novas técnicas para tornar os levantamentos mais rápidos, menos onerosos e mais quantitativos.

Atualmente têm-se um contexto de elevada disponibilidade de informação, onde se destaca a existência de inúmeras bases cartográficas e de computadores com enorme capacidade de processamento. O geoprocessamento, que é uma forma de tratamento de dados espacializados com auxílio da informática, permite executar simulações e pré-visualizações de cenários distintos sem onerar novos custos aos processos de análises. Assim, os trabalhos de levantamentos devem evoluir para atender a crescente demanda mundial por informação cada vez mais detalhada sobre solos,

compondo dados mais quantitativos, que sejam capazes de melhorar a apresentação e facilitar a interpretação das informações (BASHER, 1997). Neste caso, a cartografia digital aplicada aos solos gera muita expectativa por tratar-se de um novo desafio, com inúmeras possibilidades de avanço no que diz respeito à reunião e organização de dados de solos e meio ambiente (GALVÃO e FORMAGGIO, 2007).

O Mapeamento Digital de Solos (MDS) pode ser definido como a criação de um sistema espacial de distribuição dos solos por modelos numéricos, que permitem inferir sobre as variações espaciais e temporais dos tipos de solos e de suas propriedades, a partir das observações e conhecimentos de solos e das variáveis ambientais relacionadas (LAGACHERIE e MCBRATNEY, 2007). Este conceito já vem sendo aplicado e testado ao longo dos últimos 20 anos, e apesar de ser inquestionável o potencial do MDS, não existem parâmetros e metodologias com aplicação universal, por exemplo, quanto a densidade de amostragem, conjuntos de covariáveis ambientais, uso de dados legados, dentre outros aspectos.

Para Lagacherie (2008), o MDS se encontra em fase de amadurecimento e consolidação metodológica e apresenta algumas lacunas de conhecimento passíveis de serem exploradas com novas pesquisas. Uma primeira lacuna apontada por este autor é a necessidade de se melhorar as escalas espaciais e resoluções dos Modelos Digitais de Elevação (MDE's), que constituem a principal fonte de geração de co-variáveis utilizadas para alimentar os modelos de predição de solo (McBRATNEY et al., 2003; TEN CATEN et al., 2012). Para isto torna-se necessário avaliar diferentes bases cartográficas com maior ou menor tolerância nos processos de interpolação, visando obter melhores MDE's. Além de testar novos produtos existentes no mercado como imagens hiperespectrais de alta resolução e o mapeamento altimétrico a laser do LiDAR (Light Detection And Ranging) que é capaz de gerar dados com resolução espacial sub-métrica, que permite trabalhar em escalas ultra-detalhadas.

Outra perspectiva de pesquisa encontra-se na busca de novas co-variáveis, que podem ser obtidas pelo processamento de covariáveis simples ou através de inputs de novas fontes de dados, como os dados de levantamentos geofísicos, destacando-se a espectometria de raios gama. Neste contexto a interação com outras áreas da ciência no levantamento de

informações se torna cada vez mais necessária para operacionalizar mapeamentos cada vez mais precisos e confiáveis.

A aerogamaespectometria une métodos geofísicos e técnicas de sensoriamento remoto, sendo capaz de registrar a radiação gama emitida por rochas e solos. Foi originalmente desenvolvida para exploração de urânio (U) e tem sido utilizada para mapeamentos geológicos em área densamente vegetadas a várias décadas. Os raios gama têm comprimentos de onda em torno de 10^{-3}nm e são originados pelo decaimento dos isótopos radioativos K^{40}, Tl^{208} (Th) e Bi^{214} (U) contidos em minerais. Portanto, toda a radiação gama detectada sobre a superfície da terra é proveniente da radioatividade natural do decaimento destes três elementos: potássio (K), urânio (U) e tório (Th). E os dados obtidos pelos sensores são diretamente proporcionais a quantidade de radioelementos presentes na fonte (WILFORD e MINTY, 2007).

Um aspecto relevante nos trabalhos de mapeamento de solos que deve ser bastante estudado junto as técnicas de MDS é a amostragem. Esta é uma etapa que representa boa parte dos custos envolvidos nos trabalhos de levantamento de solos, e é uma etapa fundamental, pois é a partir dos trabalhos de campo que se verifica de fato a ocorrência dos solos em determinada região. Além disso, para que se melhore a informação de solos existente, é necessário que se colete novas amostras e ou observações de campo. No entanto, a forma de proceder esta amostragem, quanto a intensidade e localização, buscando assim contemplar a variabilidade ambiental existente de uma perspectiva quantitativa, é ainda uma questão passível de investigações.

Diante deste contexto de desafios, novas perspectivas e de uma necessidade eminente de amadurecimento metodológico das técnicas de Mapeamento Digital de Solos, o presente trabalho tem como objetivo central propor e avaliar metodologias relcionadas com a questão da amostragem e da seleção de covariáveis ambientais, visando alcançar resultados que se expressem em mapas cada vez mais precisos e úteis.

Com intuito de investigar o comportamento das técnicas de MDS abordadas neste trabalho em diferentes escalas de mapeamento, optou-se por trabalhar em áreas com distintas proporções, contemplando assim cenários locais, regionais e nacionais. Para tanto três áreas foram escolhidas como objeto de estudo neste trabalho a saber: O território do estado de Minas Gerais,

com aproximadamente 586.519,727 km^2; A área de Proteção Ambiental Sul, localizada nos domínios do Quadrilátero Ferrífero com 1.625,32 km^2; e a bacia hidrográfica do Rio Turvo Sujo, localizada na Zona da Mata mineira onde ocupa 400 km^2.

REFERÊNCIAS BIBLIOGRÁFICAS

BASHER, R. Is pedology dead and buried? Australian Journal of Soil Research, 1997, v.35, n.5, p.979-994.

CHAGAS, C.S. **Mapeamento digital de solos por correlação ambiental e redes neurais em uma bacia hidrográfica no Domínio de Mar de Morros.** Viçosa, MG: UFV, Tese (Doutorado em Solos e Nutrição de Plantas) – Universidade Federal de Viçosa. 2006; 223p.

EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA - EMBRAPA. Centro Nacional de Pesquisa de Solo. **Procedimentos normativos de levantamentos pedológicos.** / Santos, H.G. et al.. Brasília: Embrapa – SPI, 1995. 116p.

GALVÃO, L.S.; FORMAGGIO, A.R. Sensoriamento remoto hiperespectral e geração de informações pedológicas. **Boletim Informativo da Sociedade Brasileira de Ciência do Solo**, Viçosa, 2007; 32, 1, 27-31.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE. Coordenação de Recursos Naturais e Estudos Ambientais. Manual técnico de pedologia. 3ª ed. Rio de Janeiro: IBGE, 2015. 323p.

KER, J.C. Reflexões sobre levantamentos pedológicos no país. **Boletim Informativo da Sociedade Brasileira de Ciência do Solo**, Viçosa, 2007, v.32, n.1, p.13-17.

McBRATNEY, A.B. et al. On digital soil mapping. **Geoderma**, 2003; v.117, p.3-52.

MENDONÇA-SANTOS, M.L.; SANTOS, H.G. dos. The state of the art of brazilian soil mapping and prospects for digital soil mapping. In: LAGACHERIE, P. et al. Digital soil mapping: an introductory perspective. Amsterdam: Elsevier, 2007, p.39-54.

OLIVEIRA, V.A. As implicações da informatização nos levantamentos pedológicos. Boletim Informativo da Sociedade Brasileira de Ciência do Solo, Viçosa 2014; 39: 8-13.

OLIVEIRA, V.A. As implicações da informatização nos levantamentos pedológicos. **Boletim Informativo da Sociedade Brasileira de Ciência do Solo**, Viçosa, 2007; v.32, n.1, p.32-36.

RESENDE, M.; CURI, N.; REZENDE, S.B.; CORRÊA, G.F. **Pedologia:** base para distinção de ambientes.5ª ed. rev. Lavras: Editora UFLA, 2007.

ten CATEN, A.; DALMOLIN, R.S.D.; MENDONÇA-SANTOS, M.L.; GIASSON, E. Mapeamento digital de classes de solos: características da abordagem brasileira. Ciência Rural, 2012; 42:1990-1997.

UNITED STATES DEPARTMENT OF AGRICULTURE - USDA. Soil survey division. Soil conservation service. Soil Survey Staff. **Soil Survey Manual.** Rev. Enlarg. Ed. Washington, (USDA. Agriculture handbook, 18). 1993. 437p.

WILFORD, J.; MINTY, B. **The use of airborne gamma-ray imagery for mapping soils and understanding landscape processes**. Developments in Soil Science, volume 31. 2007.

CAPITULO 1

MAPEAMENTO DIGITAL DE CLASSES DE SOLO NO ESTADO DE MINAS GERAIS: AMOSTRAGEM COM HIPERCUBO LATINO CONDICIONADO BALANCEADO

RESUMO

VASCONCELOS, Bruno Nery Fernandes, D.Sc., Universidade Federal de Viçosa, setembro de 2016. **Amostragem com Hipercubo Latino Condicionado Balanceado no mapeamento digital de classes de solo no Estado de Minas Gerais.** Orientador: Elpídio Inácio Fernandes Filho. Coorientadores: João Carlos ker e Carlos Ernesto G.R. Schaefer.

A amostragem é uma questão essencial nos de levantamentos de solos, e por ser uma etapa de custo elevado, busca-se através de diferentes formas elaborar planos de amostragem que sejam mais eficientes e gerem mapas de maior exatidão. Uma proposta que vem se consagrando cada vez nos trabalhos de Mapeamento Digital de Solos (MDS) é o método de Amostragem Hipercubo Latino Condicionado - cLHS. Apesar da eficiência já comprovada deste método em relação a outros como a amostragem aleatória, o mesmo não é sensível ao problema da desproporcionalidade entre o tamanho das classes. Esta desproporcionalidade gera um desbalanceamento amostral que dificulta a detecção e ou predição de classes com menor área. Este estudo apresenta uma proposta de incremento no método de amostragem cLHS, a partir da inserção do desbalanceamento das classes de solo como um limitante (custo), afim de melhorar a precisão na predição de classes menos expressivas. O método aqui proposto foi testado no território do Estado de Minas Gerais sobre dados legados de solo (mapa de solos 1;650.000). O desempenho do método foi comparado com a amostragem Aleatória e com o método cLHS original, em cinco intensidades de amostragem diferentes 1.500, 2.000, 5.000, 10.000 e 23.500 amostras. O método proposto (cLHS balanceado) não apresentou diferença estatística dos valores de índice Kappa em relação aos outros dois métodos (Aleatório e cLHS original). Entretanto, conseguiu predizer com menores valores de erro todas as classes, mesmo em baixas intensidades de

amostragem. A melhora na classificação de classes menos expressivas evidencia o potencial do método em detectar as mesmas. Os resultados permitem concluir que o método cLHS balanceado foi mais eficiente do que os métodos Aleatório e cLHS original, na classificação de classes de menor área, sendo esta uma opção para amostragem em áreas onde ocorra expressiva desproporcionalidade entre os tamanhos de classes a serem preditas.

Palavras-chave: Random Forest, cLHS, desbalanceamento amostral.

ABSTRACT

VASCONCELOS, Bruno Nery Fernandes, D.Sc., Universidade Federal de Viçosa, September, 2016. **Digital soil classes mapping in the State of Minas Gerais: Sampling with Conditioning Latin Hypercube balanced.** Adviser: Elpídio Inácio Fernandes Filho. Co-advisers: João Carlos Ker and Carlos Ernesto G.R. Schaefer.

Sampling is a key issue in the soil survey work, and to be a costly step, is sought through various forms develop sampling plans that are more efficient and generate the most accurate maps. A sample proposal that has been devoting time in Digital Mapping works Soil (MDS) is the method of the hypercube Latino Conditioned - cLHS. Despite the proven effectiveness of this method compared to other like Random sampling, the same does not seem to be sensitive to the problem of disproportion between the size of classes. This disproportion generates a sample inbalance that hampers the detection and prediction or smaller area classes. This study presents a proposed increase in cLHS sampling method, from the insertion of the inbalance of the soil classes as a limiting (cost) in order to improve the accuracy in predicting less expressive classes. The proposed method has been tested in the State of Minas Gerais on legacy soil data (soil map 1: 650,000). The method performance was compared with the random sampling and the original cLHS method, five different sampling intensities 1.500, 2.000, 5.000, 10.000, and 23.500 samples. The proposed method (cLHS balanced) showed no statistical difference of the Kappa index values in the other two methods (Random and cLHS original). Furthermore, it could predict all classes, even at low sampling intensities with lower error

values. The improvement in less expressive class's classification demonstrates the potential of the method to detect the same. The results show that the balanced cLHS method was more efficient than the Random methods and cLHS original , the sort of smaller area classes, which is an option for sampling in areas where occur significant disproportion between the class sizes to be predicted.

Keywords: Random Forest, cLHS, inbalance sampling.

1. INTRODUÇÃO

O princípio fundamental do mapeamento digital de solos é correlacionar classes ou propriedades de solos com covariáveis ambientais que representam em maior instância os fatores de formação de solos. Para tanto empregam-se modelos numéricos ou estatísticos para inferir sobre as variações espaciais dos solos de forma mais quantitativa (LAGACHERIE e MCBRATNEY, 2007).

A amostragem é uma questão essencial nos trabalhos de levantamento de solos e tem sido cada vez mais abordada nos trabalhos de mapeamento digital de solos (TESKE et al., 2015; BAGATINI et al., 2015; BRUNGARD et al., 2015; SZATMÁRI et al., 2015; ZHANG et al., 2016; STUMPF et al., 2016; BLASCHEK e DUTTMANN, 2015). É uma etapa de custo elevado, e desta forma busca-se através de diferentes alternativas para elaborar planos de amostragem que sejam mais eficientes, propiciando melhorias na exatidão dos mapas gerados com menor custo.

Um plano de amostragem eficiente deve contemplar a variabilidade dos solos existentes na área de estudo, apresentar boa operacionalidade, o que consiste na acessibilidade dos pontos em campo, além de considerar a possibilidade de redução do número de amostras a serem coletadas a partir do conhecimento prévio dos dados legados de solo (STUMPF et al., 2016).

Segundo Brungar e Boettinger (2010) as estratégias de amostragem adotadas em levantamentos convencionais de solos não são eficientes para Mapeamento Digital de Solos (MDS), visto que existe uma seleção de locais de amostragem, subjetiva por parte dos pedólogos, para sustentar os modelos mentais de distribuição dos solos na paisagem. Desta forma a amostragem feita convencionalmente em levantamentos pedológicos é direcionada, não

aleatória e pode não representar totalmente a área de estudo do ponto de vista estatístico (HENGL, 2003).

No Mapeamento Digital de Solos (MDS) um plano de amostragem eficiente deve contemplar a variabilidade existente nas covariáveis preditivas, visto que estas representam a estratificação básica do ambiente, por meio da qual irá se predizer a distribuição espacial dos solos e seus atributos na paisagem. Uma proposta de amostragem que vem se consagrando cada vez mais nos trabalhos de MDS é a técnica denominada Hipercubo Latino condicionado (conditioned Latin Hypercube Sampling - cLHS), (BRUNGARD e BOETTINGER, 2010; ADAMCHUK et al., 2011; MENEZES et al., 2013; MULDER et al., 2012; MCKENZIE; et al., 2008; LEVI e RASMUSSEN, 2014; CARVALHO JÚNIOR et al., 2014; KIDD et al., 2015; BRUNGARD et al., 2015; Stumpf et al., 2016), a qual foi elaborada por Minasny e Mcbratney (2006). O método cLHS, promove uma amostragem aleatória estratificada que objetiva gerar uma malha amostral que contemple maximamente a variabilidade dos estratos que compõem as covariáveis ambientais utilizadas no modelo preditivo. Sendo assim, consegue-se com uma menor quantidade de amostras uma representatividade maior da variabilidade existente na área estudada (MINASNY e MCBRATNEY, 2006).

Um aspecto importante a ser considerado pelo método de amostragem quando se trabalha na predição de classes de solos, é a presença de classes de ocorrência reduzida nas áreas. Estas classes, embora menos expressivas, geralmente apresentam muita relevância do ponto de vista ambiental e de uso dos solos. Estão comumente associadas a porções importantes da paisagem como fundo de vales, encostas íngremes ou topos de vertentes em relevos estruturais.

Alguns autores tem relatado a dificuldade de predizer classes ou unidades de mapeamento (UM's) pouco expressivas em mapeamento digital de solos (TEN CATEN et al., 2012), e métodos de amostragem que tentam contemplar estas classes tem sido propostos.Teske et al. (2015) sugere uma amostragem estratificada que seja proporcional ao tamanho da área de ocorrência das classes ou UM's.

A desproporcionalidade no tamanho de área entre classes ou UM's gera um desbalanceamento da amostragem, visto que classes com maior área recebem naturalmente um número maior de amostras, sendo esta a principal

dificuldade em se treinar e validar modelos para contemplar as classes ou UM's menos expressivas. Este aspecto interfere diretamente na precisão da classificação destas classes conforme evidenciado em alguns trabalhos (HENGL et al., 2007; KIM et al., 2012; BARTHOLD et al., 2013).

A questão do desbalanceamento entre classes e amostras é considerada crucial em processos de aprendizado de máquinas e mineração de dados, pois gera efeitos negativos no desempenho dos classificadores (CHAWLA et al., 2004). No intuito de amenizar este problema, diversos trabalhos têm sido conduzidos em muitas áreas do conhecimento, principalmente relacionados à aprendizagem de máquinas (Machine Learning) (JAPKOWICZ, 2000; HAN et al., 2005; BATISTA et al., 2004; AKBANI et al., 2004; PROVOST, 2000).

Brungard et al., (2015) realizaram comparação de mapeamento digital de classes de solo em três distintas regiões dos Estados Unidos. Estes autores constataram que a diferença entre a precisão de classificação nas áreas, pode ser atribuída à existência de um desbalanceamento de amostragem, ocasionado pela desproporcionalidade no tamanho da área das classes. Desta forma, a região que tinha classes de área muito grande, abarcando cerca de 70% das amostras, foi a que apresentou maior precisão na classificação.

Neste contexto este estudo tem como principal objetivo propor e avaliar o método de amostragem cLHS balanceado, a partir da inserção da desproporcionalidade das classes de solo como um limitante (custo). A hipótese testada aqui é que ao se promover um melhor balanceamento do número de amostras por classe, o efeito da amostragem hipercubo latino condicionada irá melhorar a precisão na predição de classes menos expressivas.

2. MATERIAL E MÉTODOS

2.1. Área de estudo

O Estado de Minas Gerais está localizado no sudeste do país, com área de 586.519,727 km^2 (Figura 1).Possui heterogeneidade relacionada ao seu meio físico. Conforme a classificação climática de Koppen, o estado apresenta

um gradiente climático no sentido norte/sul, com cerca de 67 % da área sobre clima Aw (Clima tropical de savana com estação seca de inverno), predominantemente na porção norte. Na porção centro/sul do Estado predominam os tipos climáticos Cwa (Clima temperado úmido com inverno seco e verão quente) e Cwb (Clima temperado úmido com inverno seco e verão moderadamente quente) com respectivamente 21 % e 11 % da área total do mesmo (JÚNIOR, 2009).

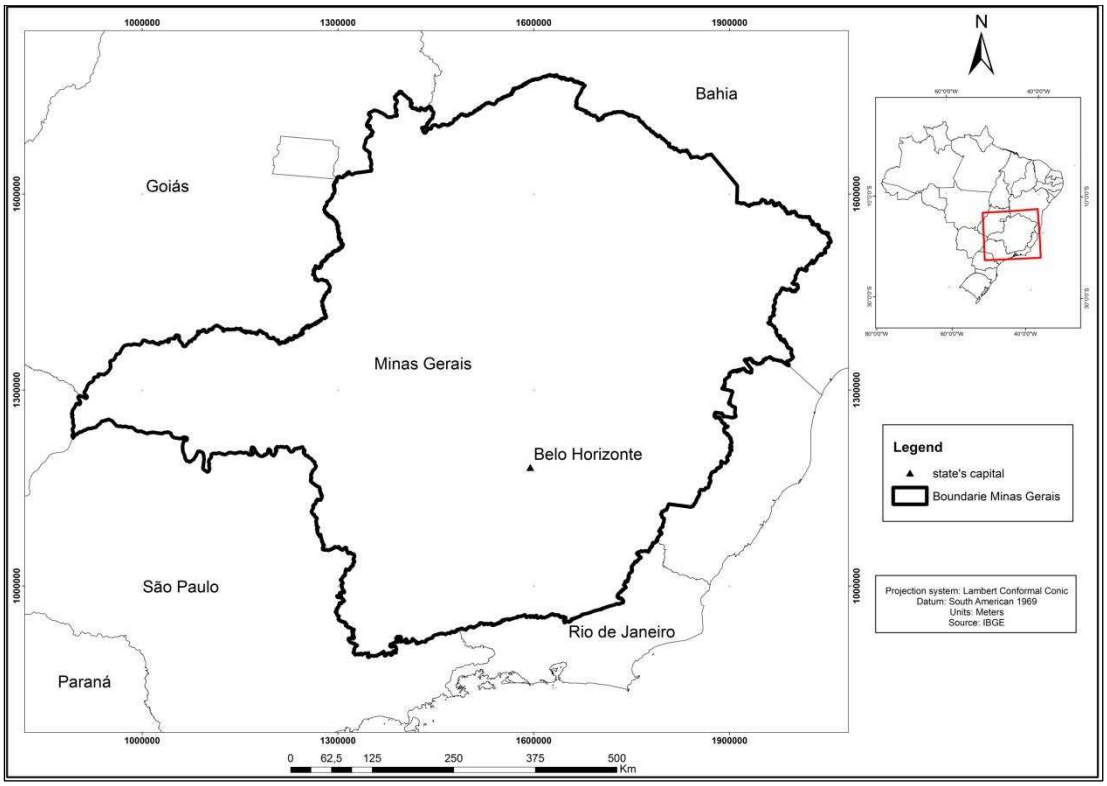


Figura 1. Localização da área de estudo.

A estrutura geológica do Estado compreende terrenos do complexo cristalino Arqueano, associados ao Cráton do São Francisco, nos quais repousa uma importante bacia sedimentar do Neoproterozóico, representada por rochas sedimentares do Grupo Bambuí. Outro grande compartimento geológico trata-se da faixa Móvel Atlântica, representada por relevos bem movimentados onde predominam rochas metamórficas intensamente dobradas e falhadas. Nesta região intercalam-se maciços montanhosos e planaltos dissecados desenvolvidos sobre um clima tropical úmido (SCHAEFER, 2013). Existe ainda uma porção no oeste do estado pertencente a Bacia Sedimentar Paleozoica do Paraná, recoberta por Basaltos da Formação Serra Geral

comumente intercalados por Arenitos da Formação Botucatu, ambos do Cretáceo.

As variabilidades climáticas e geológicas conferem uma diversidade pedológica no Estado, que é expressa pela presença de solos que contemplam a totalidade das 13 ordens do Sistema Brasileiro de Classificação de Solos.

2.2. Dados de solo

A base de dados de solos utilizada neste trabalho foi o mapa de solos na escala 1:650.000 (UFV et al, 2010) delineado convencionalmente com 303 unidades de mapeamento. Para este estudo foram consideradas as classes de solo no segundo nível categórico (subgrupo) do Sistema Brasileiro de Classificação de Solos (EMBRAPA, 2013). Foram identificadas 19 classes de solos (até 2º nível categórico) no território de Minas Gerais, sendo este o objeto classificado neste trabalho. A Figura 2 apresenta o mapa de solos utilizado com o número e a subordem das respectivas classes de solos.

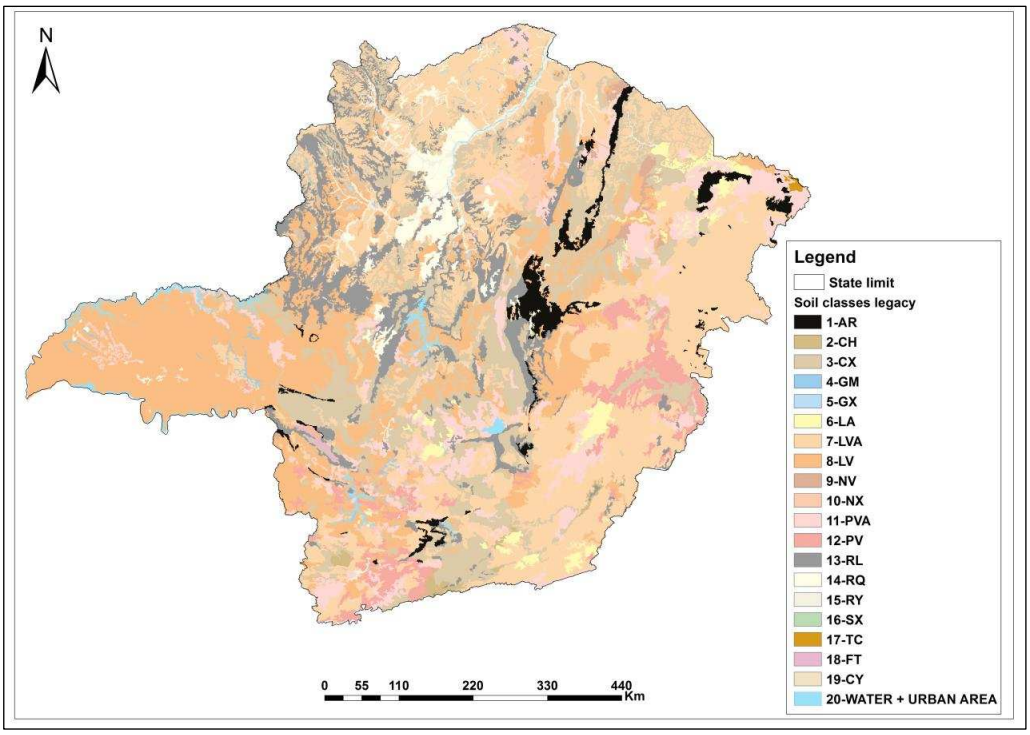


Figura 2. Mapa de solos legado do Estado de Minas Gerais – Classes: 1-AR Afloramento de Rocha; 2-CH Cambissolo Húmico; 3- CX Cambissolo Háplico; 4-GM Gleissolo Melânico; 5- Gleissolo Háplico; 6- LA Latossolo Amarelo; 7-LVA Latossolo Vermelho-Amarelo; 8-LV Latossolo Vermelho; 9-NV Nitossolo Vermelho; 10-NX Nitossolo Háplico; 11-PVA Argissolo Vermelho-Amarelo; 12-PV Argissolo Vermelho; 13-RL Neossolo Litólico; 14-RQ Neossolo Quartzarênico; 15-RY Neossolo Flúvico; 16-SX Planossolo Háplico; 17-TC Luvissolo Crômico; 18-FT Plintossolo Argilúvico; 19-CY Cambissolo Flúvico.

2.3. Covariáveis ambientais

As covariáveis ambientais utilizadas para a classificação foram selecionadas a partir de um conjunto inicial de 134 covariáveis (Tabela 1). A partir do conjunto total elaborou-se uma seleção removendo incialmente as variáveis altamente correlacionadas (correlação não linear acima de 95%) para evitar efeito de colinearidade. Em seguida buscou-se identificar dentre as variáveis categóricas aquelas que apresentassem alta similaridade. Posteriormente utilizou-se a função Importance do algoritmo Randon Forest, que ranqueia as covariáveis por ordem de importância, selecionando por fim um conjunto de 16 covariáveis apresentado na Tabela 2. As variáveis morfométricas relacionadas ao relevo foram obtidas através do software R com o pacote R-Saga, a partir do Modelo Digital de Elevação (MDE) - SRTM (Shuttle Radar Topography Mission) interpolado para a resolução de 1km (CGIAR, 2015). Além disso, utilizou-se um mapa categórico com as subdivisões dos principais compartimentos geomorfológicos do estado. Representando o fator Organismos foi utilizado o Normalized Difference Vegetation Index (NDVI), obtido a partir das imagens do sensor MODIS (Moderate Resolution Imaging Spectroradiometer) (USGS, 2016). O fator Material de Origem foi contemplado utilizando-se o Mapa Geológico (CODEMIG, 2014) e o Mapa de Geodiversidade (CPRM, 2016), além de dados aerogeofísicos de gamaespectometria com seis canais (Th, K, U, razões Th/K, Th/U e U/K), gravimetria e magnetometria nas variações: derivada vertical, intensidade total e sinal analítico (CODEMIG, 2014). As variáveis associadas ao clima foram obtidas da base de dados WorldClim – Global Climate Data (WorldClim, 2015).

Tabela 1. Conjunto total de covariáveis utilizadas, subdivididas pelos fatores de formação de solos.

Fatores de formação de solos	Covariáveis
Relief	Mapa geomorfológico – MDE -Real Surface Area - Convergence index - Cross secational curvature -Flow line curvature - General curvature - Longitudinal curvature - Maximal curvature - Minimal curvature - Plan curvature - Profile curvature - Tangencial curvature - Euclidean distance drainage - Diurnal anysotropic heating– Gradient - Hill index – Landforms - Standardized height - Mid slope position - Morphometric protection index - Normalized height -Slope - MRRTF (Multi-resolution ridge top flatness) - MRVBF (Multi-resolution valley bottom flatness) - Slope Height - Mass balance index -WTI (1) - Solar radiation diffuse 1 - Solar radiation diffuse 2 - Solar radiation direct 1 -Solar radiation direct 2 - Solar radiation duration 1 - Solar radiation duration 2 - Solar radiation total 1 -Solar radiation total 2 - Surface specific points - Terrain Ruggednes Index - Terrais Surface convexity - Topographic Position Index - Topographic Wetness Index (TWI) - Valley Index – Valley - Valley Depth - Vector Ruggedness Measure
Climate	Tmax, Tmim, Tmean, Preciptation for the 12 months, and 19 Bioclimatic variables
Parent Material	Geological and Geodversity maps (1:1.000.000), Concentration of the K, U, Th elements - Ratio between the Th / K, Th/U, U/K (Gammaspectrometry), Magnetometry (full strengh, vertical derivative, analytical signal), Gravimetry
Organism	NDVI (Normalized Difference Vegetation Index)

Tabela 2. Covariáveis selecionadas, subdivididas pelos fatores de formação de solos

Fatores de formação de solos	Covariáveis
Relevo	Mapa geomorfológico - MRRTF (Multi-resolution ridge top flatness) - MRVBF (Multi-resolution valley bottom flatness) - Solar radiation diffuse 1
Clima	Temperatura mínima no mês 7 – Bio 4 (Sazonalidade de Temperatura) – Bio 13 (Precipitação no mês mais úmido) - Bio 18 (Precipitação no trimestre mais quente) – Bio 19 (Precipitação no trimestre mais frio)
Material de Origem	- Mapas Geológico (1:1.000.000) – Mapa de Geodiversidade (litologias)(1:1.000.000) - Concentração do elemento K (gamaespectometria) – Razão entre concentrações dos elementos Th/K
Posição espacial	Latitude - Longitude

2.4. Proposta metodológica de coleta de amostras para treinamento

A extração das amostras das classes de solo foi feita de forma computacional, no software R, com 10 repetições sobre o mapa de solos existente (legacy data) para o estado (UFV et al., 2010), a partir de três métodos de amostragem: aleatório, cLHS e cLHS balanceado. Destaca-se que o termo amostragem aqui empregado, refere-se à coleta de amostras para treinamento e validação do classificador. A densidade de amostragem, apesar de ser fundamental, é uma questão ainda não bem definida para trabalhos de MDS (BAGATINI et al., 2015; CARVALHO JÚNIOR et al., 2014). Existem algumas propostas como a de Zhu (2000) que sugere adotar-se um número de amostras 30 vezes o número de classes a serem preditas. Neste estudo adotou-se como referência, para o estabelecimento do número de amostras de treinamento, a recomendação do número de observações sugerida pelos dois documentos que tratam sobre normativas de levantamentos pedológicos no Brasil (EMBRAPA, 1995; IBGE, 2015). Desta forma adotou-se uma escala final de 1:500.000 (calculada a partir do tamanho do pixel 1.000 m, e da acuidade visual do ser humano 0,002 m), caracterizando um mapeamento de reconhecimento de baixa intensidade. Neste tipo de mapeamento recomenda-se amostrar um perfil completo e uma amostra extra a cada 856 km^2, gerando cerca de 1.500 amostras para a área de estudo. Já quanto a densidade de observações recomenda-se 0,04 observações por km^2, gerando um total de 23.500 amostras. Foram adotadas cinco diferentes intensidades de amostragem, tendo como limite inferior 1.500 amostras e superior 23.500 amostras além de valores intermediários de 2.000, 5.000 e 10.000 amostras, sendo estes valores aplicados em cada um dos três métodos de amostragem testados.

2.4.1. Amostragem aleatória

A amostragem aleatória representa o esquema de amostragem mais simplificado que é comumente empregado, devido ao fato de eliminar a subjetividade e por apresentar fácil reprodutibilidade. Além disso, este tipo de amostragem foi tomado neste trabalho como a testemunha, permitindo assim um parâmetro comparativo para os demais métodos, dentro da mesma base de

dados. As malhas de pontos aleatórios foram geradas a partir do comando spsample no software R (R CORE TEAM, 2016).

2.4.2. Amostragem Hipercubo Latino Condicionado (cLHS)

O método de amostragem cLHS tem seus fundamentos no método LHS (Latin Hypercube Sampling) que segue a idéia de um quadrado latino onde existe somente uma amostra em cada linha e em cada coluna, porém com uma generalização deste conceito para um número arbitrário de dimensões (MINASNY e MCBRATNEY, 2006). Desta forma este método tem como ideia central representar a variabilidade espacial das covariáveis a partir de um conjunto de amostras distintas. Para tanto o cLHS subdivide cada covariável em estratos igualmente prováveis que corresponde ao tamanho do conjunto da amostra, e objetiva com um número pré-estabelecido de pontos amostrais, contemplar maximamente a variabilidade expressa por estes estratos. O método foi aplicado, utilizando-se o conjunto de covariáveis previamente selecionadas (Tabela 2), às cinco intensidades de amostragem com 100.000 interações. O número de interações representa as tentativas de experimentação que o método realiza buscando a melhor conformação de amostragem. Foi utlizada a biblioteca clhs (ROUDIER et al., 2012) no software R (R CORE TEAM, 2016).

2.4.3. Amostragem cLHS balanceado

A proposta metodológica apresentada neste trabalho consiste em uma alteração do método cLHS, a partir da utilização de um custo (cost) obtido pelo cálculo de proporcionalidade entre as áreas das classes. Para tanto se considerou os seguintes pressupostos descritos (Tabela 3).

- O número de amostras por classe foi estabelecido a partir da área de cada classe, considerando-se uma densidade de amostragem de uma amostra a cada 100 km^2.
- O desbalanceamento entre o número de amostras das classes foi estabelecido em 1:10. Desta forma qualquer uma das classes menores tem pelo menos 1 amostra a cada 10 amostras presentes em classes maiores.

Tabela 3. Dadosutilizada para gerar o método cLHS balanceado

CLASSES DE SOLOS		Área (Km²)	Número de amostras	Desbalanceamento
1-AR	AFLORAMENTO DE ROCHA	12692,00	159	0,10
2-CH	CAMBISSOLO HUMICO	2195,48	159	0,10
3-CX	CAMBISSOLO HAPLICO	89812,90	898	0,56
4-GM	GLEISSOLO MELANICO	1533,59	159	0,10
5-GX	GLEISSOLO HAPLICO	581,71	159	0,10
6-LA	LATOSSOLO AMARELO	8699,93	159	0,10
7-LVA	LATOSSOLO VERMELHO AMARELO	158966,00	1589	0,90
8-LV	LATOSSOLO VERMELHO	120026,00	1200	0,76
9-NV	NITOSSOLO VERMELHO	2173,58	159	0,10
10-NX	NITOSSOLO HAPLICO	3079,02	159	0,10
11-PVA	ARGISSOLO VERMELHO AMARELO	40648,90	406	0,26
12-PV	ARGISSOLO VERMELHO	19247,90	192	0,12
13-RL	NEOSSOLO LITOLICO	45287,40	452	0,28
14-RQ	NEOSSOLO QUARTZARENICO	11679,20	159	0,10
15-RY	NEOSSOLO FLUVICO	10053,90	159	0,10
16-SX	PLANOSSOLO HAPLICO	46,63	159	0,10
17-TC	LUVISSOLO CROMICO	341,38	159	0,10
18FT	PLINTOSSOLO ARGILUVICO	583,59	159	0,10
19-CY	CAMBISSOLO FLUVICO	391,38	159	0,10
20	AGUA + AREA URBANA	5300,95	159	0,10

– Calculou-se a proporção de amostragem que foi obtida a partir da divisão entre o maior número de amostras, evidentemente associado à classe de maior área, e o número mínimo de amostras que está associado às classes que tem as menores áreas.

– -Por fim utilizou-se a proporcionalidade expressa na coluna proporção daTabela 3, para gerar um arquivo raster que foi introduzido como custo no

método cLHS. Desta forma, classes que tem maior área tem maior custo amostral, por outro lado as classes de menor área apresentam menor custo amostral. A presença deste custo faz com que o cLHS ao experimentar células de baixo custo, mantenha estas amostras dentre as escolhidas, e,ao amostrar células de alto custo, tenda a excluir esta amostra do conjunto selecionado. Por fim as classes de menor área recebem um incremento amostral em detrimento das de maior área.

2.5. Avaliação

A avaliação do trabalho consistiu na comparação entre os resultados obtidos para os três métodos de amostragem, frente as cinco diferentes intensidades de amostragem. A exatidão dos mapas foi avaliada através da matriz de erro, onde se tem nas colunas os dados de referência e nas linhas os dados classificados, sendo a diagonal principal a representação do nível de concordância entre ambos os mapas (CONGALTON e GREEN, 1999). Para compensar os acertos ao acaso que não são levados em consideração pela matriz de erro optou-se por utilizar o Índice Kappa, que trata-se de uma estatística multivariada discreta utilizada para medir a concordância entre dados estimados e os de referência, mas que desconsidera os acertos ocorridos ao acaso. O Índice Kappa normalmente varia entre 0 e 1, sendo 0 a ausência de concordância e 1 concordância total (CONGALTON e GREEN, 1999).

Além da avaliação do índice Kappa para cada um dos métodos, foi avaliado também o erro de classificação de cada uma das classes, com vistas a analisar melhor o comportamento dos métodos a nível de individualidade das classes. Parta tanto foi calculado o desbalanceamento amostral de cada classe, e a desproporcionalidade em relação ao número total de amostras. O desbalanceamento reflete a diferença do número de amostras entre a maior classe e as demais. Já a desproporcionalidade reflete a diferença entre o número de amostras presente nas classes em relação ao número total de amostras (1.500, 2.000, 5.000, 10.000 e 23.500).

3. RESULTADOS

Na Figura 3 é possível observar a variabilidade dos valores do índice Kappa em cada uma das 10 repetições realizadas para cada método nas diferentes intensidades de amostragem, evidenciando assim a importância de se fazer repetições quando se está utilizando métodos de amostragem fundamentados na aleatoriedade. Observa-se também que o método cLHS balanceado é o que apresenta maior variabilidade dentre as repetições.

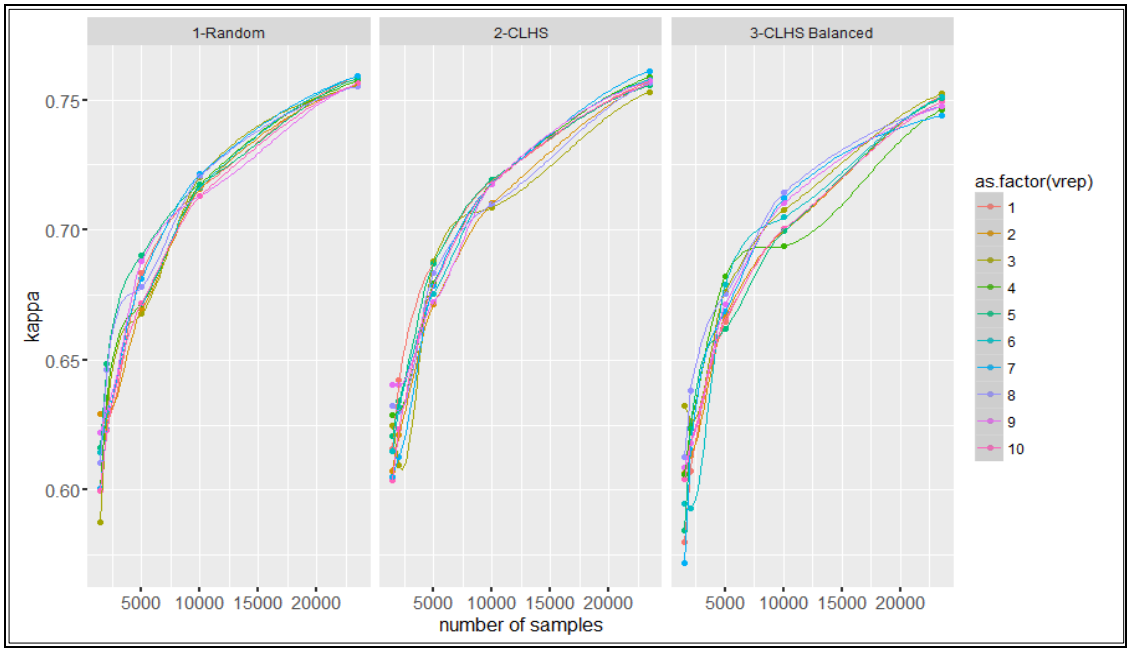


Figura 3. Variabilidade dos valores do índice kappa entre as 10 repetições para os três métodos de amostragem avaliados.

Na Figura 4 apresenta a média das 10 repetições dos valores do índice kappa para os três métodos de amostragem variando com a intensidade de amostragem. Todas as curvas têm um comportamento semelhante e exponencial, evidenciando que existe um ganho mais expressivo na precisão da classificação até próximo de 10.000 pontos amostrais. A partir deste ponto observa-se menores incrementos nos valores de kappa a medida que se aumenta o número de amostras.

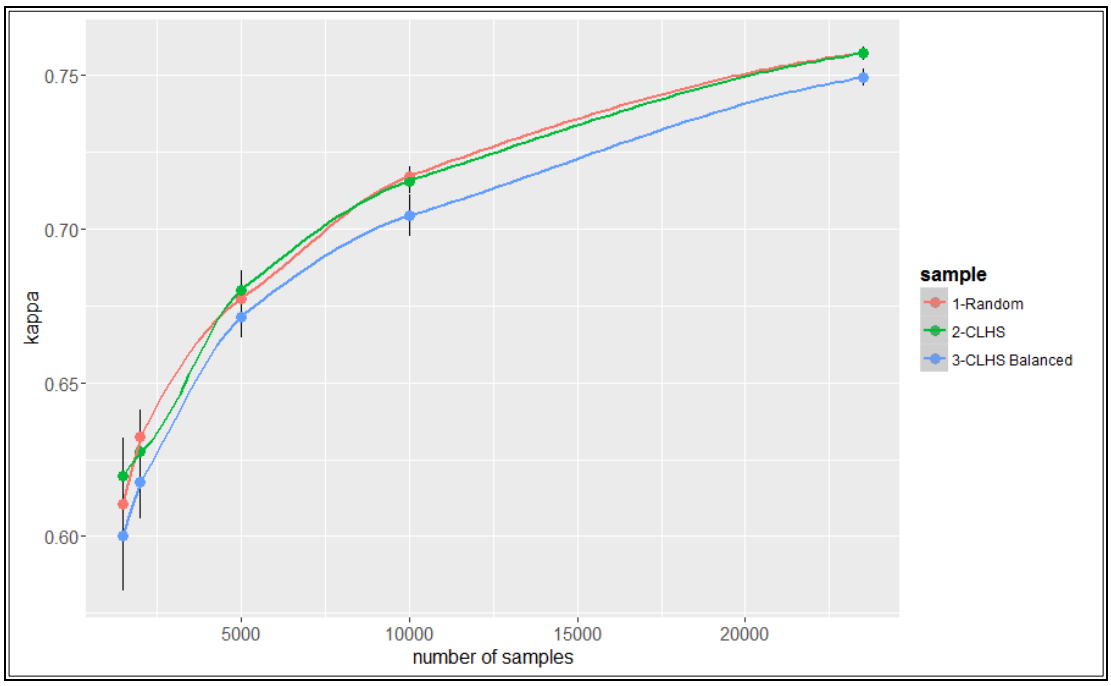


Figura 4. Média dos valores do índice kappa obtida das 10 repetições para os três métodos de amostragem avaliados nas diferentes intensidades de amostragem.

Os maiores valores de kappa estão associados aos métodos cLHS e aleatório. No entanto o método cLHS balanceado, que apresentou os menores valores do índice Kappa, não apresentou diferença estatística dos demais métodos pelo Teste Z a 5 % de probabilidade.

Quando se avalia o erro na classificação de cada classe individualmente, observa-se que o método cLHS balanceado apresenta valores de erro mais baixos para praticamente todas as classes (Figura 5 e 6), apresentando comportamento inverso a este somente nas classes de maior área (7 e 8), onde todos os métodos tiveram valores de erro baixos. Isso evidencia que mesmo que este método tenha apresentado valores ligeiramente menores de índice kappa, foi mais eficiente na classificação das classes de menor área que comumente apresentam erros maiores devido a baixa densidade de amostragem atribuída as mesmas.

A partir dos resultados obtidos foi possível observar um incremento gradativo na eficácia dos métodos avaliados com relação à classificação das classes de menor área, com o aumento do número de amostras utilizadas. A Figura 6 demonstra o número de classes não classificadas em cada um dos métodos de amostragem. O método Aleatório foi o que apresentou maior perda de classes, principalmente nas de menor intensidade de amostragem. Já o

método cLHS só não conseguiu classificar uma das classes, quando se adotou 1.500 e 2.000 amostras. O método cLHS balanceado se apresentou como o método mais eficiente, pois classificou todas as classes mesmo nas menores intensidades de amostragem.

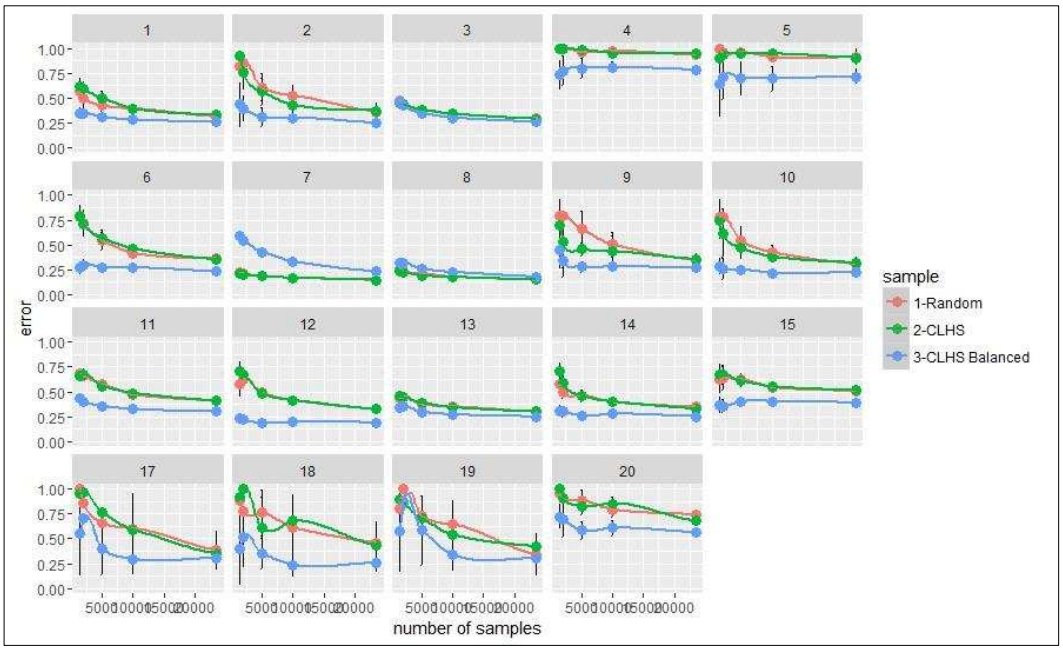
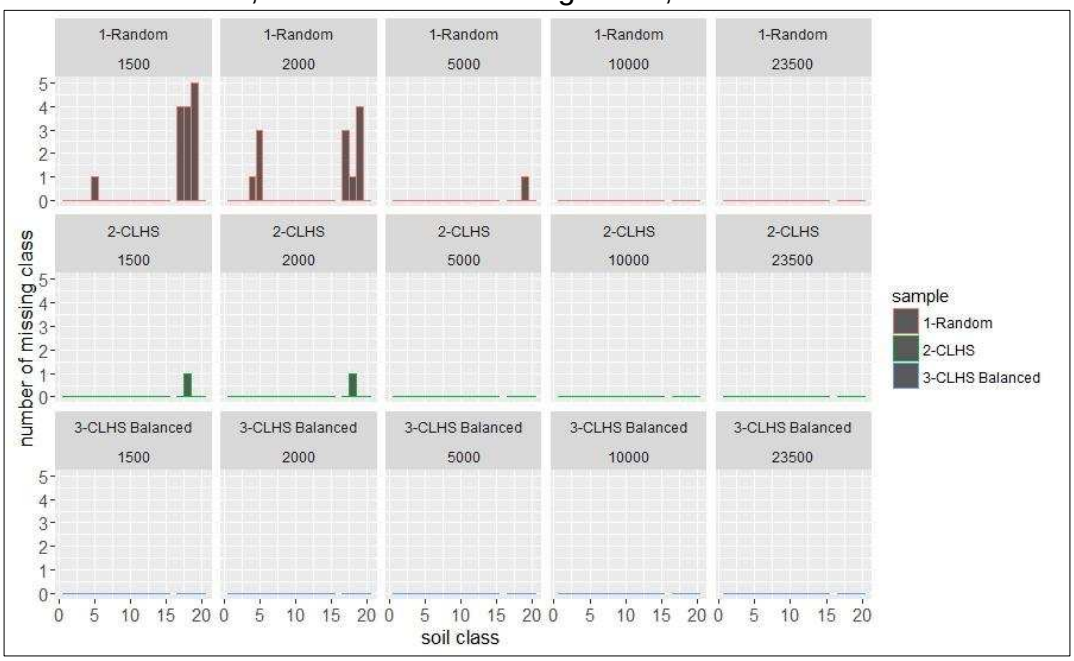


Figura 5. Erros associados a cada classe na classificação feita com as amostras obtidas para cada um dos métodos de amostragem. Classes: 1-AR Afloramento de Rocha; 2-CH Cambissolo Húmico; 3- CX Cambissolo Háplico; 4-GM Gleissolo Melânico; 5- Gleissolo Háplico; 6- LA Latossolo Amarelo; 7-LVA Latossolo Vermelho-Amarelo; 8-LV Latossolo Vermelho; 9-NV Nitossolo Vermelho; 10-NX Nitossolo Háplico; 11-PVA Argissolo Vermelho-Amarelo; 12-PV Argissolo Vermelho; 13-RL Neossolo Litólico; 14-RQ Neossolo Quartzarênico; 15-RY Neossolo Flúvico; 16-SX Planossolo Háplico; 17-TC Luvissolo Crômico; 18-FT Plintossolo Argilúvico; 19-CY Cambissolo Flúvico.



Figrua 6. Número de classes não classificadas por cada um dos métodos nas diferentes intensidades de amostragem.

A identificação de que o método cLHS apresenta desbalanceamento na elaboração da malha amostral, e que a estratégia de balancear esta amostragem é justificável é evidenciada na Figura 7, onde é possível identificar que o desbalanceamento de amostragem nas classes é sempre maior no método cLHS, mesmo quando o número de amostras é baixo. A única classe em que o método apresenta um comportamento diferente é a de número 7. Nesta classe, que apresenta a maior área dentre todas as classes, o desbalanceamento só diminui no método cLHS balanceado próximo do valor máximo de 23.500 amostras.

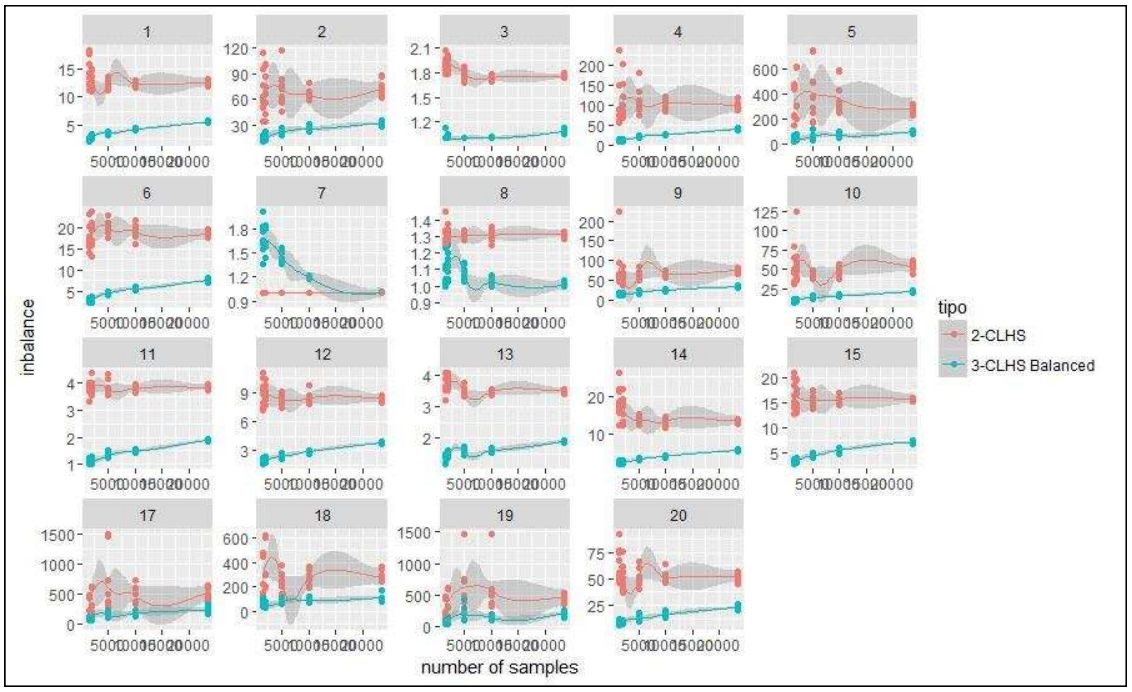


Figura 7. Desbalanceamento de amostragem em cada uma das classes entre os métodos cLHS e cLHS balanceado. Classes: 1-AR Afloramento de Rocha; 2-CH Cambissolo Húmico; 3- CX Cambissolo Háplico; 4-GM Gleissolo Melânico; 5- Gleissolo Háplico; 6- LA Latossolo Amarelo; 7-LVA Latossolo Vermelho-Amarelo; 8-LV Latossolo Vermelho; 9-NV Nitossolo Vermelho; 10-NX Nitossolo Háplico; 11-PVA Argissolo Vermelho-Amarelo; 12-PV Argissolo Vermelho; 13-RL Neossolo Litólico; 14-RQ Neossolo Quartzarênico; 15-RY Neossolo Flúvico; 16-SX Planossolo Háplico; 17-TC Luvissolo Crômico; 18-FT Plintossolo Argilúvico; 19-CY Cambissolo Flúvico.

Na Figura 8 que apresenta uma relação entre o erro e o desbalanceamento é possível observar a relação direta que existe entre o desbalanceamento de amostras e o maior erro na classificação das classes. Novamente a classe 7 apresenta comportamento diferente, visto que o erro é maior no método cLHS balanceado.

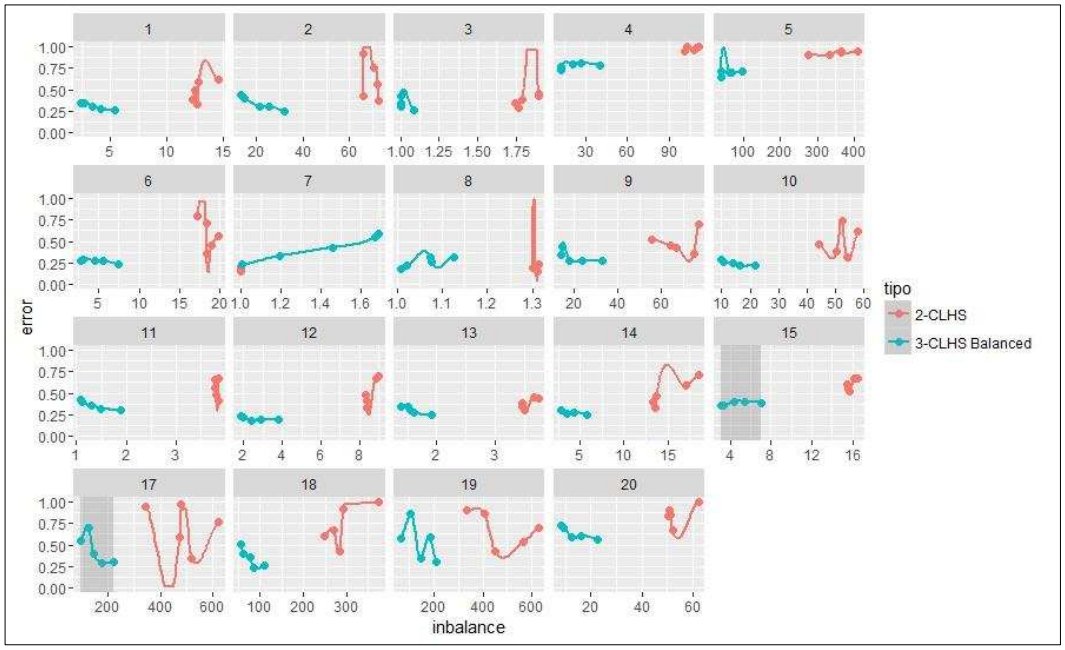


Figura 8. Relação entre desbalanceamento e erro em cada uma das classes.

4. DISCUSSÃO

Os três métodos de amostragem comparados não apresentaram diferença estatística nos valores de kappa pelo teste Z a 5 %. O método de amostragem cLHS apresentou valores de índice Kappa muito semelhantes aos encontrados pelo método de amostragem Aleatório. Este fato se torna mais evidente a medida que se aumenta o número de amostras. Por outro lado, pode considerar-se que o primeiro (cLHS) demonstrou superioridade em relação ao segundo, visto que mesmo com valores semelhantes de índice Kappa, o método cLHS deixou de classificar somente uma classe, enquanto o método aleatório perdeu 4, 5 e 1 classes, nas menores intensidades de amostragem 1.500, 2.000 e 5.000 respectivamente. Embora não tenha sido comparado de forma idêntica, visto que neste trabalho não se avaliou o comportamento do método cLHS em relação a variabilidade presente nas covariáveis ambientais, o melhor desempenho do método cLHS em relação ao Aleatório é concordante com resultados obtidos em outros trabalhos (CARVALHO JÚNIOR et al., 2014; MINASNY e MCBRATNEY, 2007).

A não detecção destas classes está associada a escala de trabalho adotada neste trabalho (1:500.000), estabelecida a partir da resolução espacial dos dados (1000 metros), bem como pelo pequeno número de amostras utilizadas para treinamento, e consequentemente prejudicando a sua predição.

Apesar do método cLHS apresentar um bom desempenho, foi possível diagnosticar neste estudo que este método resulta em uma amostragem com desbalanceamento amostral elevado entre as classes. Além disso, parece que, embora alguns trabalhos já tenham constatado a dificuldade em predizer classes de pouca expressividade (HENGL et al., 2007; KIM et al., 2012; BARTHOLD et al., 2013; TESKE et al., 2015), este aspecto do desbalanceamento amostral ainda é pouco discutido na literatura científica relacionada a mapeamento digital de solos.

O método de amostragem cLHS é, portanto, um método eficiente mas passível de adequações que possam melhorar ainda mais seu desempenho diante de condições específicas. Neste sentido, a proposta apresentada neste trabalho vem de encontro com alguns outros trabalhos que avaliam não só o desempenho, mas também o incremento de modificações no método. Stumpf et al. (2016), propuseram um método de amostragem cLHS modificado a partir da incorporação de dados legados de solo juntamente com aspectos de acessibilidade dos locais em campo, afim de gerar um conjunto de pontos amostrais alternativos dentro do conjunto gerado pelo cLHS original, possibilitando assim que os autores pudessem escolher posteriormente pontos operacionalmente mais viáveis de serem coletados em campo.

O incremento do desbalanceamento das classes como um custo no método cLHS, gerando o cLHS balanceado, demonstrou-se eficiente para amenizar o problema de classificação de classes menos expressivas em termos de área, tanto pelo fato de que nenhuma classe foi desprezada quanto pelo fato de que o erro associado a classificação destas classes foi sempre menor do que aquele obtido para o método cLHS original. Além disso, não houve diferença estatística entre os valores de kappa deste método em relação aos demais. Sendo assim, manteve-se a exatidão do mapa gerado com incremento de qualidade na classificação de classes de menor área.

As classes consideradas menos expressivas, devido à área de ocorrência restrita, representam solos importantes do ponto de vista de uso e ocupação. A classe 19 (Cambissolo Flúvico) por exemplo, que não foi detectada pelo método aleatório e pelo cLHS original, com 1.500 e 2.000 amostras, representa uma área nobre do ponto de vista de uso. São solos associados a terraços fluviais onde predominam condições de fertilidade natural melhores e relevo favorável a mecanização. As classes 4 (Gleissolo

Melânico) e 5 (Gleissolo Háplico), que não foram detectadas pelo método de amostragem Aleatória, são muito relevantes do ponto de vista ambiental, pois representam áreas de preservação ambiental. Além disso, estão diretamente associadas com a dinâmica hídrica, em nível regional, e com o acúmulo de carbono devido às condições de anaerobiose em que permanecem constantemente.

Outra classe não detectada pelo método Aleatório foi a classe 17 (Luvissolo Crômico), que apresenta-se no mapa original como uma pequena área no extremo nordeste do Estado, e caracteriza-se por apresentar boa fertilidade natural, sendo utilizada predominantemente para pastagens e ou lavouras. Este aspecto evidencia a dificuldade de se individualizar classes taxonomicamente semelhantes, e que muitas vezes não possuem um estratificador ambiental que possa ser usado para separá-las.

Neste contexto um caminho promissor que pode representar um real avanço na qualidade dos mapas de solos, através do MDS, é a inserção de mapas de propriedades dos solos para distinguir classes que não são ambientalmente separáveis. No caso dos Luvissolos e dos Argissolos, o principal fator de distinção entre estas classes, é a atividade da argila, sendo alta (>= 27 cmolc/kg) para os primeiros, e menor (< 27 cmolc/kg) para os segundos (EMBRAPA, 2013). Sendo assim, a elaboração de um mapa do atributo diagnóstico atividade da argila, gerado a partir de uma modelagem geoestatística, por exemplo, poderia ser um valioso plano de informação para ajudar a distinguir estas classes.

Outro exemplo desta possibilidade refere-se a separação entre Latossolos com textura média de Neossolos Quartzarênicos (UFV, 2010), na região noroeste do Estado (margem esquerda do rio São Franciso). Nesta região ambos os solos ocorrem em áreas de relevo plano a suave ondulado, dificultando assim a distinção destas duas classes a partir do relevo. As variações litológicas também não são bons estratificadores, pois muitas vezes tratam-se de sedimento quaternários de granulometria variável (CODEMIG, 2014), e com distribuição espacial totalmente aleatória. Neste caso o emprego de uma mapa de argila, que evidenciasse em algum nível regiões com maior conteúdo desta fração, poderiam ser muito úteis no delimeamento dos limites entre estas duas classes nesta área do Estado.

5. CONCLUSÕES

- A análise de eficiência das classificações feitas em MDS, somente pela avaliação dos valores do índice kappa, pode não ser suficiente para detectar problemas relacionados a classes de menor área. Sendo assim, é desejável que se faça alguma avaliação de erro em cada classe individualmente.
- A inserção do desbalanceamento das classes como um custo no método de amostragem cLHS, demonstrou ser uma estratégia eficiente para melhorar a predição de classes menos expressivas em termos de área de ocorrência no mapeamento digital de solos.
- Os valores de índice kappa obtidos para o método aqui proposto cLHS balanceado, não diferiram estatisticamente dos demais métodos. Além disso, o erro atribuído a cada classe foi menor neste método. Sendo assim, considera-se que o método cLHS balanceado apresentou desempenho superior ao cLHS original e ao método Aleatório, no mapeamento digital de classes no Estado de Minas Gerais.
- A utilização de métodos de amostragem que tenham capacidade de predizer melhor classes menos expressivas em termos de área de ocorrência, é justificável pelo fato de que, além destas classes apresentarem relevância ambiental e de uso do solo, são classes comumente detectadas e mapeadas em trabalhos de levantamento convencional.
- Uma possibilidade de avanço na qualidade dos mapas de solo a partir do MDS, pode ser a utilização de mapas de propriedades dos solos como fonte de informação no processo de classificação, para individualizar classes taxonomicamente semelhantes, que não apresentem bons estratificadores ambientais

6. REFERÊNCIAS BIBLIOGRÁFICAS

ADAMCHUK, V.I. et al. Using targeted sampling to process multivariate soil sensing data. **Geoderma**, v. 163, n.1-2, p.63–73, 2011.

AKBANI, R.; KWEK, S.; JAPKOWICZ, N. Applying Support Vector Machines to Imbalanced Datasets. **Lnai**, v.3201, p.39–50, 2004.

BAGATINI, T.; GIASSON, E.; TESKE, R. Seleção de densidade de amostragem com base em dados de áreas já mapeadas para treinamento de modelos de árvore de decisão no mapeamento digital de solos. **Revista Brasileira de Ciencia do Solo**, v.39, n.4, p.960–967, 2015.

BARTHOLD, F. K. et al. Land use and climate control the spatial distribution of soil types in the grasslands of Inner Mongolia. **Journal of Arid Environments**, v.88, p.194–205, 2013.

BATISTA, G.E.A.P.A.; PRATI, R.C.; MONARD, M.C. A study of the behavior of several methods for balancing machine learning training data. **ACM SIGKDD Explorations Newsletter**, v.6, n.1, p.20, 2004.

BLASCHEK, M.; DUTTMANN, R. A stratified two-stage sampling design for digital soil mapping in a Mediterranean basin. v.17, p.6408, 2015.

BREIMAN, L. Random Forests. **Machine learning**, v.45.1, p.5–32, 2001.

BRUNGARD, C.; BOETTINGER, J.L. Conditioned Latin Hypercube Sampling: Optimal Sample Size for Digital Soil Mapping of Arid Rangelands in Utah, USA. **Media**, n. December, 2010.

BRUNGARD, C.W. et al. Machine learning for predicting soil classes in three semi-arid landscapes. **Geoderma**, v. 239, p. 68–83, 2015.

CGIAR. Consortium for Spatial Information. SRTM 90m Digital Elevation Data. Available at:<http://srtm.csi.cgiar.org/>. Accessed Fev 2015.

CHAWLA, N.V; JAPKOWICZ, N.; DRIVE, P. Editorial : Special Issue on Learning from Imbalanced Data Sets. **ACM SIGKDD Explorations Newsletter**, v.6, n.1, p.1-6, 2004.

CODEMIG - COMPANHIA DESENVOLVIMENTO ECONÔMICO DE MINAS GERAIS– MAPA GEOLÓGICO (www.portaldageologia.com.br) acesso em 26/05/2016. 2014.

CONGALTON, R.G.; GREEN, K. **Assessing the accuracy of remotely sensed data: principles and practices**. New York: Lewis Publishers, 1999, 160p.

CPRM–COMPANHIA DE PESQUISA DE RECURSOS MINERAIS - Mapa Geodversidade (www.cprm.gov.br/geobank) acesso em 26/05/2016. 2016.

CARVALHO JÚNIOR, W. et al. Método do hipercubo latino condicionado para a amostragem de solos na presença de covariáveis ambientais visando o mapeamento digital de solos. **Revista Brasileira de Ciencia do Solo**, v.38, n.2, p.386–396, 2014.

EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA (EMBRAPA). Centro Nacional de Pesquisa de Solo. **Sistema Brasileiro de Classificação de solos**. 2013. 3 ed. Rio de Janeiro, 353p.

EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA (EMBRAPA). Centro Nacional de Pesquisa de Solo. **Procedimentos normativos de levantamentos pedológicos.** / Santos, H.G. et al.. Brasília: Embrapa – SPI, 1995. 116p.

HAN, H.; WANG, W.; MAO, B. Borderline-SMOTE : A New Over-Sampling Method in. p.878–887, 2005.

HENGL, T. et al. Methods to interpolate soil categorical variables from profile observations: Lessons from Iran. **Geoderma**, v.140, n.4, p.417–427, 2007.

HENGL, T.; ROSSITER, D.G. & STEIN, A. Soil sampling strategies for spatial prediction by correlation with auxiliary maps. Aust. J. Soil Res., 41:1403-1422, 2003.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE. www.ibge.gov.br acesso em 10/06/2016. 2016.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE. Coordenação de Recursos Naturais e Estudos Ambientais. **Manual técnico de pedologia.** 3ª ed. Rio de Janeiro: IBGE, 2015. 430 p.

JAPKOWICZ, N. The Class Imbalance Problem: Significance and Strategies. **Proceedings of the 2000 International Conference on Artificial Intelligence**, p. 111–117, 2000.

JÚNIOR, A. DE S. Aplicação Da Classificação De Köppen Para O Zoneamento Climático Do Estado De Minas Gerais. p.101, 2009.

KIDD, D. et al. Operational sampling challenges to digital soil mapping in Tasmania, Australia. **Geoderma Regional**, v.4, p.1–10, 2015.

KIM, J. et al. Multi-scale Modeling of Soil Series Using Remote Sensing in a Wetland Ecosystem. **Soil Science Society of America Journal**, v.0, n.0, p.0, 2012.

LAGACHERIE, P., MCBRATNEY, A.B., Chapter 1. Spatial soil information systems and spatial soil inference systems: perspectives for Digital Soil Mapping. In: P. Lagacherie, A.B. McBratney and M. Voltz (Editors), **Digital Soil Mapping: An Introductory Perspective**. 2007.

LEVI, M.R.; RASMUSSEN, C. Covariate selection with iterative principal component analysis for predicting physical soil properties. **Geoderma**, v.219-220, p.46–57, 2014.

MCKENZIE, N.J.; et al. **Guidelines for surveying soil and land resources**. [s.l: s.n.].

MENEZES, M. D. DE et al. Digital Soil Mapping Approach on Fuzzy Logic and Field Expert Knowledge. **Ciência e Agrotecnologia**, v.37, n.4, p.287–298, 2013.

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

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

MULDER, V.L.; DE BRUIN, S.; SCHAEPMAN, M.E. Representing major soil variability at regional scale by constrained Latin Hypercube Sampling of remote sensing data. **International Journal of Applied Earth Observation and Geoinformation**, v. 21, n. 1, p. 301–310, 2012.

PROVOST, F. Machine Learning from Imbalanced Data Sets 101 Extended Abstract. **Proceedings of the AAAI'2000 workshop on imbalanced data sets**, 2000.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. 2016.

ROUDIER, P.; HEWITT, A.E. & BEAUDETTE, D.E. A conditioned latin hypercube sampling algorithm incorporating operational constraints. In: MINASNY, B.; MALONE, B.P. & McBRATNEY, A.B., eds. Digital soil assessments and beyond. London, CRC Press/Balkema, 2012. p.227-232.

SCHAEFER, C.E.G.R. Bases físicas da Paisagem Brasileira: Estrutura Geológica, Relevo e Solos. **Tópicos em Ciência do Solo VIII.** Sociedade Brasileira de Ciência do Solo. Viçosa-MG. 2013.

STUMPF, F. et al. Incorporating limited field operability and legacy soil samples in a hypercube sampling design for digital soil mapping. **Journal of Plant Nutrition and Soil Science**, n. June, 2016.

SZATMÁRI, G.; BARTA, K.; PÁSZTOR, L. An application of a spatial simulated annealing sampling optimization algorithm to support digital soil mapping. v.64,p. 35–48, 2015.

TEN CATEN, A. et al. Mapeamento digital de classes de solos : características da abordagem brasileira. **Ciência Rural**, v.42, p.1989–1997, 2012.

TESKE, R.; GIASSON, E.; BAGATINI, T. Comparação De Esquemas De Amostragem Para Treinamento De Modelos Preditores No Mapeamento Digital De Classes De Solos. **Revista Brasileira de Ciência do Solo**, v.39, n.1, p.14–20, 2015.

Universidade Federal de Viçosa (UFV); Fundação Centro Tecnológico de Minas Gerais (CETEC-MG); Universidade Federal de Lavras (UFLA); Fundação Estadual do Meio Ambiente (FEAM). **Mapa de Solos Do Estado de Minas Gerais: legenda expandida.** Belo Horizonte: Fundação Estadual do Meio Ambiente, 2010.

WorldClim – Global Climate Data. Disponível em (<http://www. worldclim.com.br). Acesso em maio de 2015.

ZHANG, S.-J. et al. An heuristic uncertainty directed field sampling design for digital soil mapping. **Geoderma**, v. 267, p.123-136, 2016.

ZHU, A. X. **Mapping soil landscape as spatial continua: The neural network approachWater Resources Research**, 2000.

CAPÍTULO 2

AVALIAÇÃO DE DADOS LEGADOS NO BALANCEAMENTO DE AMOSTRAGEM COM HIPERCUBO LATINO CONDICIONADO PARA MAPEAMENTO DIGITAL DE CLASSES DE SOLO

RESUMO

VASCONCELOS, Bruno Nery Fernandes, D.Sc., Universidade Federal de Viçosa, Septembro, 2016. **Avaliação de dados legados no balanceamento de amostragem com hipercubo latino condicionado para mapeamento digital de classes de solo.** Orientador: Elpídio Inácio Fernandes Filho. Coorietadores: João Carlos Ker and Carlos Ernesto G.R. Schaefer.

A utilização de dados legados de solo em mapeamento digital de solos (MDS) é uma forma real de avançar no conhecimento da distribuição dos solos. A possibilidade de melhoria desta informação, através de técnicas de Mapeamento Digital de Solos, pressupõem a execução de novos trabalhos de amostragem. Um aspecto já identificado em alguns trabalhos trata da dificuldade em se amostrar e predizer determinadas classes que, apesar de possuírem relevância ambiental, ocorrem em áreas consideravelmente restritas. Desta forma estas classes são desfavorecidas por planos de amostragem que acabam por não contemplar as mesmas. Este estudo objetivou avaliar a utilização de um mapa legado de solos, na confecção de um plano de amostragem balanceado através do uso do método de amostragem Hipercubo Latino Condicionado (cLHS balanceado), comparando o mesmo com outros dois métodos, Aleatório e cLHS, e na predição das classes de solo, a qual foi realizada por três classificadores: Random Forests (RF), Gradient Boosting Machine (GBM) e C5.0. Para tanto foram geradas três malhas amostrais de 100 pontos cada, onde foram identificados em campo a classe de solo de ocorrência. Como treinamento foi utilizada uma malha amostral de 100 pontos e a validação foi efetuada com um conjunto externo de pontos composto pelos pontos das duas outras malhas não utilizadas no treinamento (200 pontos). Os resultados evidenciam o potencial de uso da metodologia de balanceamento, no entanto a eficiência do mesmo está intimamente relacionada com a

informação existente no dado legado. Os melhores valores de ínidce Kappa foram obtidos para o método clhs balanceado com os classificadores GBM e RF, e o pior valor deste índice resultou do método cLHS com o classificador RF.

Palavras-chave: Random Forest, DSM, dados legados, GBM, C5.0.

ABSTRACT

VASCONCELOS, Bruno Nery Fernandes, D.Sc., Universidade Federal de Viçosa, September, 2016. **Legacy data evaluation in sampling balancing with hypercube latino conditioning for digital mapping of soil classes.**Adviser: Elpídio Inácio Fernandes Filho. Co-advisers: João Carlos Ker and Carlos Ernesto G. R. Schaefer.

The use of soil legacy data in digital soil mapping (MDS) is a real way to advance the knowledge of soil distribution. The possibility of improving this information through Digital Mapping techniques Solos presuppose the implementation of new sampling work. One aspect identified already in some studies is the difficulty involved in sampling and predict certain classes that despite their environmental relevance occur at considerably restricted areas. Thus these classes are disadvantaged by sampling plans that end up not look the same. This study aimed to evaluate the use of a legacy of soil map, in making a balanced sampling plan by using the sampling method Latin Hypercube Conditioned (balanced cLHS) comparing the same with the other two methods, Random and cLHS, and prediction of soil types, which was performed by three classifiers: Random Forests (RF) Boosting Gradient Machine (GBM) and C5.0. Therefore, we generated three sample meshes of 100 points each, which were identified in the field the occurrence of soil class. As training we used a sampling grid of 100 points and validation was performed with an external set of points consisting of the points of the other two meshes not used in training (200 points). The results show the potential use of balanced methodology, however its efficiency is closely related to the information in the legacy map. The best values of Kappa index were obtained for the balanced

clhs method classifiers GBM and RF, and the worst value of this index resulted from cLHS method with the RF classifier.

Keywords: Random Forest, DSM, legacy data, GBM, C5.0.

1. INTRODUÇÃO

Atualmente existe uma necessidade eminente de se unir conhecimentos já existentes aos mecanismos tecnológicos cada vez mais eficientes e promissores. Na ciência do solo este contexto pode ser exemplificado com a questão dos levantamentos e mapeamentos de solos. Através de décadas estes trabalhos foram executados por profissionais que geraram um novo ramo na Ciência do Solo denominada Pedometria. No entanto novos desafios têm surgido frente aos desenvolvimentos tecnológicos demandando assim que os pedólogos busquem por meio de pesquisas, adoção de novas técnicas afim de tornar os levantamentos de solos mais rápidos e mais quantitativos, conforme ressaltado por Chagas (2006).

Neste contexto o Mapeamento Digital de Solos (MDS), que pode ser definido como a criação de um sistema espacial de distribuição dos solos por modelos numéricos, que permitam inferir sobre as variações espaciais e temporais dos tipos de solos, bem como de suas propriedades, a partir das observações e conhecimentos de solos, vinculados às covariáveis ambientais relacionadas (LAGACHERIE e MCBRATNEY, 2007).

Uma questão relevante no mapeamento digital de solos (MDS) é o uso dos dados de legados (CARRÉ et al., 2007), sendo o reaproveitamento destes dados uma possibilidade real de redução nos custos econômicos, aliada ao incremento de conhecimento sobre a distribuição dos solos nas áreas de trabalho. Além de propiciarem uma valorização de trabalhos já desenvolvidos em levantamentos de solo anteriores.

Estes dados são derivados de levantamentos tradicionais de solo, elaborados a partir de interpretações das relações existentes entre solos e paisagem que se expressam em modelos mentais elaborados pelos pedólogos (CARRÉ et al., 2007). No entanto a utilização dos dados legados pode representar um desafio visto que muitas vezes existe uma desuniformidade dos mesmos, associada com erros de localização e georreferenciamento, falta de

harmonia e imprecisão na descrição de perfis e indisponibilidade de dados numéricos (LAGACHERIE & MCBRATNEY, 2007).

Diante da existência dos dados legados e das possibilidades apresentadas pelo MDS surgem algumas questões como: qual a necessidade de se gerar novos mapas onde já existe informação de solos?E qual a capacidade de melhoria destes mapas a partir do emprego das novas técnicas? Quanto a necessidade de se melhorar tal informação Dalmolin e ten Caten (2015) justificam que o uso cada vez mais intensivo e a ocupação desordenada do solo implicam em degradação do mesmo, e que existe uma carência de informação mais detalhada sobre este recurso, visando assim melhorar as ações de planejamento.

Com relação à capacidade de se melhorar a informação existente, têm-se por exemplo a desagregação e harmonização de polígonos proposta por Odgers et al., (2014) que visa gerar novos mapas, a partir dos mapas legados, com unidades de mapeamento com somente uma classe de solo, desmembrando assim aquelas que anteriormente eram constituídas de uma associação entre duas ou mais classes. No entanto para esta ou outra abordagem que vise melhorar a informação já existente é necessário que se colete novos dados, cruciais para o estabelecimento de planos de amostragem eficientes em promover esta melhoria.

Um plano de amostragem eficiente deve contemplar a variabilidade ambiental expressa nas covariáveis preditivas, por meio da qual irá se predizer a distribuição espacial dos solos na paisagem. O método do Hipercubo Latino condicionado (conditioned Latin Hypercube Sampling - cLHS) foi elaborado por Minasny e Mcbratney (2006) e busca estabelecer uma amostragem aleatória estratificada que objetiva gerar uma malha amostral que contemple maximamente a variabilidade dos estratos que compõem as covariáveis ambientais utilizadas.

A desproporcionalidade no tamanho de área entre classes ou UM's gera um desbalanceamento da amostragem, que dificulta a predição das classes ou UM’s de menor expressividade, conforme já relatao por alguns autores como ten Caten et al. (2012) e Teske et al. (2015). Além disso, Chawla et al. (2004) ressaltam que o desbalanceamento entre classes e amostras é considerada crucial em processos de aprendizado de máquinas e mineração de dados, pois gera efeitos negativos no desempenho dos classificadores.

O reaproveitamento de dados legados representa uma possibilidade real de minimizar a quantidade de novas amostras a serem coletadas, reduzindo assim custos no incremento de conhecimento sobre a distribuição dos solos. Os dois principais tipos de dados legados de solo que devem ser distinguidos são os mapas de solo existentes bem como os perfis de solo já coletados e ou pontos de observações, que podem ser utilizados como covariáveis preditoras ou como conjunto de validação (LAGACHERIE e McBRATNEY, 2007). Com relação ao uso de mapas de solos existentes, Pahlavan-rad et al. (2016) destacam duas principais abordagens de MDS com estes dados: a primeira trata-se do uso destes mapas para direcionar a amostragem de solos em campo; e a segunda utiliza os mapas como covariáveis preditoras ou para derivar outras covariáveis. Dentre estas possibilidades de uso dos mapas legados, diversos trabalhos têm sido realizados com emprego dos mesmos (KEMPEN et al., 2009; PÁSZTOR et al., 2010; NKWUNONWO, 2015; VAYSSE E LAGACHERIE, 2015; TESKE et al., 2015; ZHANG et al., 2016; MOSLEH et al., 2016).

O presente trabalho tem como objetivo principal avaliar a utilização do mapa de solos legado da bacia rio Turvo Sujo, localizada na região da Zona da Mata mineira, para gerar uma proposta de amostragem através do Hipercubo Latino Condicionado, que seja balanceada pela desproporcionalidade existente no tamanho das áreas das unidades de mapeamento que compõem o mapa legado.

2. MATERIAL E MÉTODOS

2.1. Área de estudo

A bacia do rio Turvo Sujo está localizada na Zona da Mata mineira entre as coordenadas 42º 40’ e 43º 00’ longitude oeste e 20º 39’ e 20º 55’ de latitude sul. É uma sub-bacia da bacia hidrográfica do Rio Doce e compreende uma área de aproximadamente 400 km^2 (Figura 1).

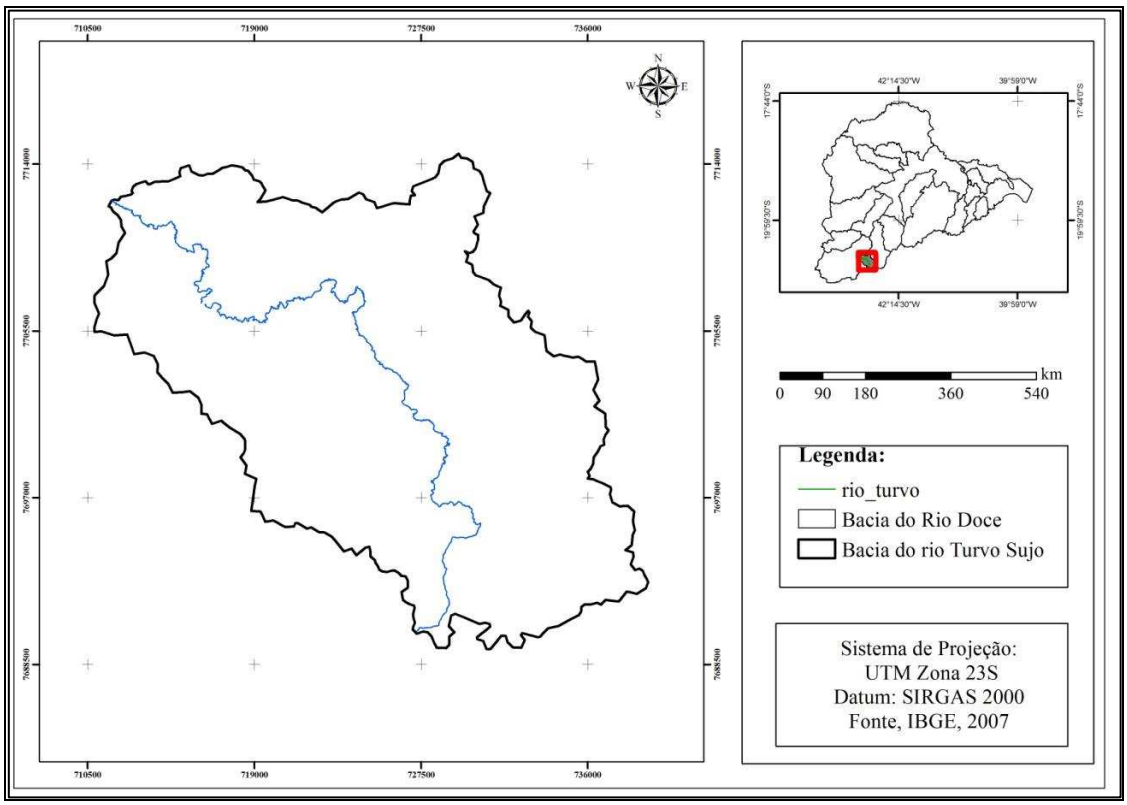


Figura 1. Localização da área de estudo.

A bacia apresenta cotas médias de 775 m em uma paisagem tipicamente representativa do domínio morfoclimático dos mares de morros. O relevo é predominantemente forte ondulado com encostas de perfil convexo-concâvo entremadas de vales colmatados por sedimentos caracterizando assim terraços e leito maiores (CORRÊA, 1984). A geologia regional é constituída basicamente de Ortognaisses do Complexo Piedade, datado do Paleoproterozoico, com presença frequente de rochas máficas oriundas de diques. Pedologicamente trata-se de um domínio de Latossolos, sendo os Vermelho-Amarelos os mais comuns. Ocorrem também Cambissolos Háplicos, via de regra associados a encontas côncavas-côncavas. Os terraços são frequentemente ocupados por Argissolos Vermelho-Amarelos, podendo também ocorrer Cambissolos Flúvicos. E nas porções mais baixas da paisagem predominam os Gleissolos Háplicos, juntamente com Neossolos Flúvicos, estes últimos mais presentes na calha do rio Turvo Sujo propriamente dito.

O clima da região é do tipo Cwa classificação de Köppen, temperado quente, com estação seca de abril a setembro e chuvosa de outubro a março, e precipitação média anual de 1200 mm.

2.2. Dados legados de solo

Foi utilizado um mapa de solos na escala 1:100.000 (MELO, 2009). O mapa foi delineado convencionalmente com 5 unidades de mapeamento a saber: PVAd (Argissolo Vermelho Amarelo Distrófico típico + Cambissolo Háplico Tb distrófico); GXbd (Gleissolo Háplico Tb distrófico + Neossolo Flúvico Tb distrófico + Argissolo Vermelho Amarelo Tb distrófico); CXbd (Cambissolo Háplico Tb distrófico); LVA1 (Latossolo Vermelho-Amarelo Distrófico típico + Latossolo Vermelho Distrófico típico); LVA2 (Latossolo Vermelho-Amarelo Distrófico típico + Cambissolo Háplico Tb distrófico).

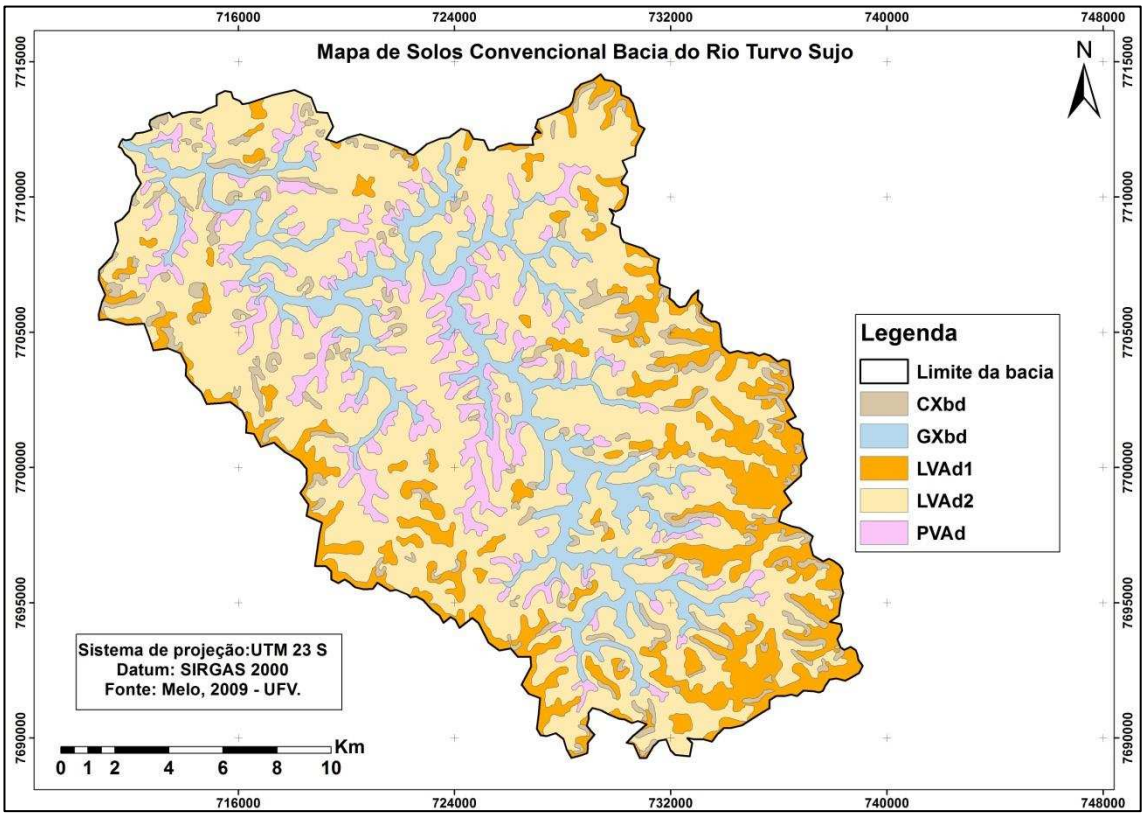


Figura 2. Mapa de solos legado da bacia hidrográfica do Rio Turvo Sujo. Unidades de Mapeamento: CXbd Cambissolo Háplico Tb distrófico; GXbd Gleissolo Háplico Tb distrófico + Argissolo Vermelho-Amarelo Distrófico típico + Neossolo Flúvico Tb distrófico; LVAd1 Latossolo Vermelho-Amarelo Distrófico típico + Latossolo Vermelho Distrófico típico; LVAd2 Latossolo Vermelho-Amarelo Distrófico típico + Cambissolo Háplico Tb distrófico

2.3. Covariáveis ambientais

As covariáveis morfométricas (Tabela 1) foram obtidas através do software R com o pacote R-Saga, a partir do Modelo Digital de Elevação (MDE) do sensor PALSAR do satélite ALOS com resolução de 12,5 m disponibilizadas

pela Agência Japonesa Exploração Aeroespacial (JAXA) no site global.jaxa.jp. Além disso, foi gerado um mapa geomorfológico com os principais compartimentos da paisagem estratificados em: Topos planos, encostas côncavas e convexas, terraços e leitos maiores. Este mapa foi gerado também a partir dos dados altimétricos da mesma imagem supracitada.

Tabela 1. Covariáveis morfométricas utilizadas

Covariáveis Morfométricas (12,5 m resolução espacial)	
Covariáveis	**Breve Descrição**
Faces de exposição (Aspect)	Orientação em relação aos pontos cardiais de cada face do relevo
Índice de Convergência (Convergence Index)	Calcula um índice de convergência / divergência em relação ao escoamento superficial
Curvatura de seção transversal (Curvature cross secational)	
Curvatura Plana (Plan curvature)	Descreve a curvatura em uma seção horizontal na encosta
Curvatura de Perfil (Profile curvature)	Descreve o segundo mecanismo de acumulação
Índice de Balanço de Massa (Mass. Balance Index)	Índice que representa o balanço de massa em cada pixel
Modelo Digital de Elevação (MDE)	Representa a elevação em cada célula do modelo
Posição média da Encosta (Mid Slope Position)	Representa a distância em relação ao topo e vale, variando entre 0 e 1. Cobre as regiões mais quentes da declividade
Elevação Normalizada (Normalized Height)	É uma medida de altura relativa de um ponto ao invés de seu valor elevação propriamente dito. Atribui o valor 1 para o ponto mais alto e o valor 0 para a posição mais baixa dentro de uma respectiva área de busca
Área Real da Superfície (Real Surface Area)	Não considera a superfície como projetada, e sim a superfície real
Declividade (Slope)	Representa a declividade local que pode ser angular ou em porcentagem
Elevação da declividade (Slope Height)	
Solrad difuse1	Radiação solar difusa incidente no mês de janeiro
Solrad direct1	Radiação solar direta incidente no mês de janeiro
Convexidade da superfície terrestre (Terrain Surface Convexity)	É calculado como a razão entre o número de células que têm curvatura positiva (células convexas) para o número de todas as células válidas dentro de um raio de pesquisa especifico
Textura da superfície terrestre (Terrain Surface Texture)	Divide a textura da superfície em 8, 12 ou 16 classes
Profundidade dos vales (Valley Depth)	Inverte a elevação, deriva as redes de drenagem e calcula a distância vertical até os mesmos
Vale (valley)	
Índice de Umidade Topográfica (WTI1)	Descreve a tendência de cada célula em acumular água

2.4. Dados de solos e metodologias de amostragem

Os dados de solos foram obtidos a partir de campanhas de campo onde se percorreru a bacia coletando pontos de observação georreferenciados com a informação da classe de solo existente nos mesmos.

Foram estabelecidas três malhas de amostragem pelos seguintes métodos: Aleatório, cLHS e cLHS balanceado. Cada uma das três malhas continha 100 pontos, portanto foram coletados um total de 300 pontos na área de estudo, perfazendo assim uma densidade de amostragem de aproximadamente 1 ponto a cada 130 ha.

2.4.1. Amostragem aleatória

A malha de pontos aleatórios foi gerada a partir do comando spsample no software R (R CORE TEAM, 2016). Esta amostragem é comumente empregada, devido ao fato de eliminar a subjetividade e por apresentar fácil reprodutibilidade. Desta forma adotou-se a mesma como a testemunha, permitindo assim um parâmetro comparativo para os demais métodos, dentro da mesma base de dados.

2.4.2. Amostragem Hipercubo Latino Condicionado (cLHS)

O método de amostragem cLHS tem seus fundamentos no método LHS (Latin Hypercube Sampling) que segue a idéia de um quadrado latino onde existe somente uma amostra em cada linha e em cada coluna, porém com uma generalização deste conceito para um número arbitrário de dimensões (MINASNY e MCBRATNEY, 2006). Desta forma este método tem como idéia central representar a variabilidade espacial das covariáveis a partir de um conjunto de amostras distintas. Para tanto o cLHS subdivide cada covariável em estratos igualmente prováveis que corresponde ao tamanho do conjunto da amostra, e objetiva com um número pré-estabelecido de pontos amostrais, contemplar maximamente a variabilidade expressa por estes estratos. O método foi utilizado nas covariáveis apresentadas na Tabela 1 com 100.000 interações, sendo que o número de interações representa as tentativas de experimentação que o método realiza buscando a melhor conformação de

amostragem. Foi utlizada a biblioteca clhs (Roudier et al., 2012) no software R (R CORE TEAM, 2016).

2.4.3. Amostragem cLHS balanceado

O método de amostragem do Hipercubo Latino Condicionado balanceado (cLHS balanceado) trata de uma alteração do método cLHS, a partir da utilização de um custo (cost) obtido pelo cálculo de proporcionalidade entre as áreas das classes. Para tanto se considerou os seguintes pressupostos descritos (Tabela 2).

Tabela 2. Tabela utilizada para gerar o método cLHS balanceado.

Unidades de mapeamento		Área (ha)	Número de amostras	Proporção
1-PVAd	Argissolo Vermelho-Amarelo	3528,85	22	0,14
2-GXbd	Gleissolo Háplico + Argissolo Vermelho Amarelo + Neossolo Flúvico	4511,80	28	0,18
3-CXbd	Cambissolo Háplico	2303,11	14	0,09
4-LVAd1	Latossolo Vermelho Amarelo + Latossolo Vermelho	6290,34	39	0,26
5-LVAd2	Latossolo Vermelho Amarelo + Cambissolo Háplico	23768,39	148	1,00

- O número de amostras por classe foi estabelecido a partir da área de cada classe, considerando-se uma densidade de amostragem de uma amostra a cada 100 ha.
- O desbalanceamento entre o número de amostras das classes foi estabelecido em 1:10. Desta forma qualquer uma das classes menores tem pelo menos 1 amostra a cada 10 amostras presentes em classes maiores.
- Calculou-se a proporção de amostragem que foi obtida a partir da divisão entre o maior número de amostras, evidentemente associado à classe de maior área, e o número mínimo de amostras que está associado às classes que tem as menores áreas.
- Por fim utilizou-se a proporcionalidade expressa na coluna proporção do Tabela 2, para gerar um arquivo raster que foi introduzido como custo no

método cLHS. Desta forma, classes que tem maior área tem maior custo amostral, por outro lado as classes de menor área apresentam menor custo amostral. A presença deste custo faz com que o cLHS ao experimentar células de baixo custo, mantenha estas amostras dentre as escolhidas, e ao amostrar células de alto custo, aumentando assim o custo amostral, tenda a excluir esta amostra do pacote selecionado. Por fim as classes de menor área recebem um incremento amostral em detrimento das de maior área.

2.5. Treinamento e Validação

Utilizou-se o conjunto total de amostras (300), particionado sempre em dois conjunto, um de treinamento e outro de validção. Uma das malhas foi utilizada como conjunto de treinamento e as outras duas como conjunto de validação. Desta forma treinava-se com 100 pontos e validava com 200. A exatidão dos mapas foi avaliada através da matriz de erro, ou de confusão, onde se tem nas colunas os dados de referência e nas linhas os dados classificados, sendo a diagonal principal a representação do nível de concordância entre ambos os mapas (CONGALTON e GREEN, 1999). Para compensar os acertos ao acaso que não são levados em consideração pela matriz de erro optou-se por utilizar o Índice Kappa que trata-se de uma estatística multivariada discreta utilizada para medir a concordância entre dados estimados e de referência, mas que desconsidera os acertos ocorridos ao acaso. O Índice Kappa normalmente varia entre 0 e 1, sendo 0 ausência de concordância e 1 concordância total (CONGALTON e GREEN, 1999).

2.6. Modelos preditivos

A predição dos mapas foi realizada no software R (R CORE TEAM, 2016) utilizando-se o comando predict da biblioteca caret. Neste trabalho foram avaliados três classificadores diferentes: Random Forest (RF); Gradient boosting machines (GBM); e o C5.0. O RF foi desenvolvido como uma extensão dos modelos de Árvores de classificação. Trata-se de uma combinação de árvores preditoras que dependem dos valores de um vetor aleatório amostrado independentemente, mas com a mesma distribuição para todas as árvores (BREIMAN, 2001). Dentre as vantagens deste método

destaca-se que ele trabalha conjuntamente com variáveis categóricas e numéricas, apresenta grande eficiência e rapidez no processo de treinamento e possui três parâmetros a serem ajustados, sendo estes o número de variáveis por nó (mtry), o número de árvores do modelo (ntree) e o número de nós na ramificação final das árvores, sendo adotados nas predições os parâmetros default do método.

O princípio básico do Gradiente Boosting Machine (GBM) parte de uma dada função de perda e um classificador pouco eficiente como, por exemplo, as árvores de regressão, em seguida o algoritmo procura encontrar um modelo que minimize a função de perda. O GBM normalmente inicia com a melhor estimativa da resposta (por exemplo, a média da resposta da regressão). O gradiente (ou erro residual) é calculado, e um modelo é ajustado para os resíduos minimizando assim a função de perda. O atual modelo é adicionado ao modelo anterior, e o procedimento se repete sucessivamente por um número especificado pelo usuário (número de interações) Khun e Jhonson (2013).

O terceiro classificador testado foi o C5.0, que é uma versão mais avançada do modelo de classificação C4.5 que tem funcionalidades adicionais, tais como aumentar e os custos desiguais para diferentes tipos de erros. O modelo tem cria uma única árvore de classificação, apresentando melhorias, em relação ao seu antecessor, pois cria árvores menores. Por exemplo, o algoritmo vai combinar condições de não ocorrência dividindo-as em várias categorias, além de realizar um procedimento de poda final global que são tentativas para remover as sub-árvores através de uma abordagem custo-complexidade. As sub-árvores serão retiradas até atingir a taxa de erro superior ao erro padrão estabelecido para não ocorrer mais poda. Khun e Jhonson (2013) afirman que tais procedimentos adicionais tendem a criar árvores mais simples do que o algoritmo anterior.

Desta forma foram preditos nove mapas provenientes dos três métodos de amostragem em cada um dos três classificadores.

4. RESULTADOS

Um primeiro aspecto que pode ser analisado a partir dos resultados obtidos é a freqüência de distribuição das amostras entre as cinco classes

preditas. A Tabela 3 evidencia que o método de amostragem cLHS apresentou a distribuição de amostras mais desbalanceada, concentrando 85 das 100 amostra na classe dominante (LVAd). O método aleatório contemplou todas as classes existentes e apresentou um desbalanceamento intermediário. Já o método cLHS balanceado apresentou uma distribuição amostral um pouco mais equilibrada do que os demais, principalmente em relação às classes CXbd e PVAd, sendo este um aspecto satisfatório. Por outro lado este método (cLHS balanceado) não alocou nenhuma de suas amostras na classe RYbd.

Tabela 3. Número de amostras alocadas em cada uma das classes pelos métodos de amostragem utilizados

Método de amostragem	**Número de amostras por classe (total 100 amostras)**				
	CXbd	**GXbd**	**LVAd**	**PVAd**	**RYbd**
Aleatório	5	3	77	13	2
cLHS	5	3	85	6	1
cLHS balanceado	12	3	68	17	0

Portanto a partir da distribuição amostral obtida, fica evidente a dificuldade de ambos os métodos alocarem amostras nas classes de menor área, particularmente GXbd e RYbd.

As Figuras 3, 4 e 5 apresentam os valores de exatidão global (Accuracy) e do índice Kappa obtidos para os três métodos de amostragem, por cada um dos classificadores, RF, GBM e C5.0 respectivamente.

Em todas as figuras pode-se notar elevados valores de Accuracy, evidenciando assim o efeito da presença de uma classe muito expressiva, que acarreta um alto acerto ao acaso por parte do classificador. Quando analisados os valores do índice Kappa nota-se uma diferença significativa nos três classificadores.

Os melhores valores do índice Kappa foram obtidos para o método do cLHS balanceado a partir dos três classificadores testados. Já o método cLHS apresentou os piores resultados nos três classificadores, sendo o melhor de seus resultados obtido pelo classificador C5.0. Com o método aleatório

obtiveram-se resultados intermediários, sempre melhores do que os obtidos com o cLHS, e sempre piores aos encontrados com cLHS balanceado.

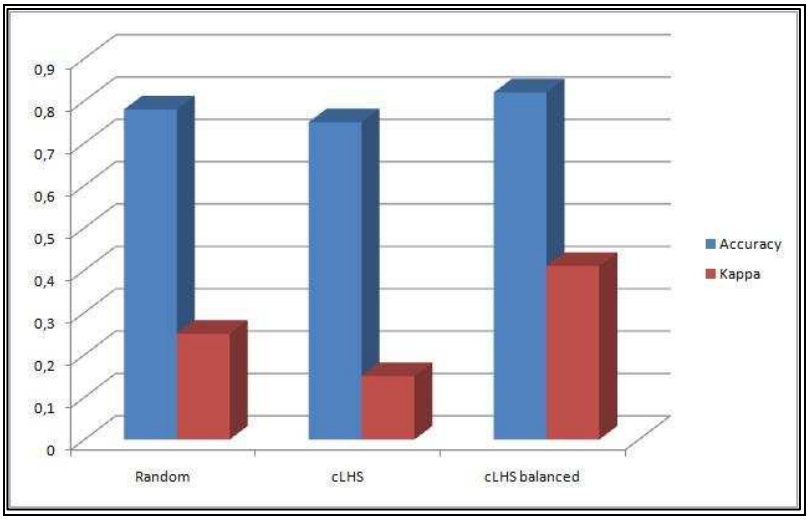


Figura 3. Valores do índice Kappa e do Accuracy (exatidão global) para os três métodos de amostragem com o classificador Random Forest.

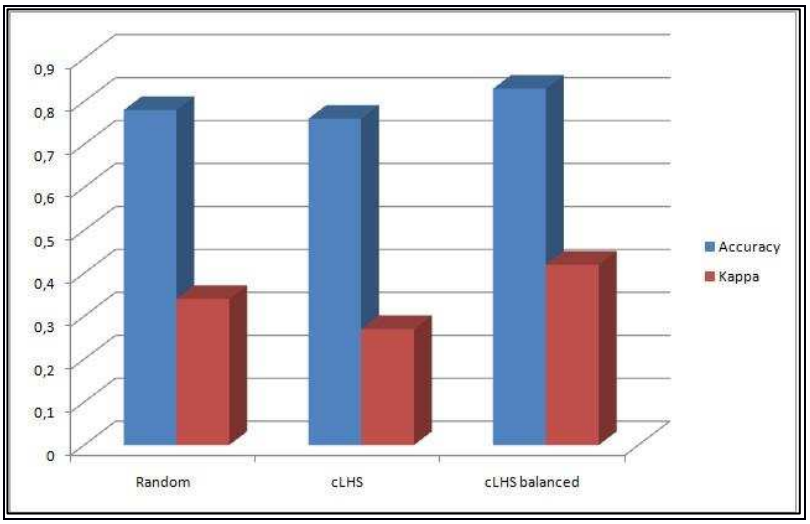


Figura 4. Valores do índice Kappa e do Accuracy (exatidão global) para os três métodos de amostragem com o classificador Gradient Boosting Machine (GBM).

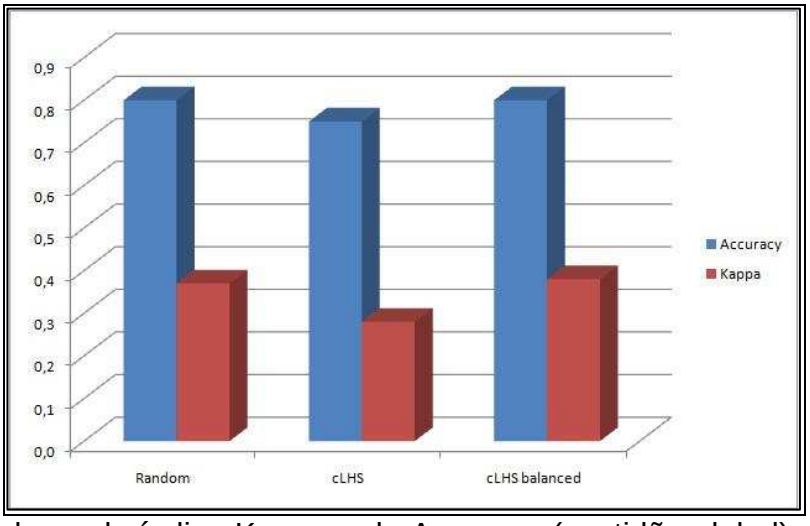


Figura 5. Valores do índice Kappa e do Accuracy (exatidão global) para os três métodos de amostragem com o classificador C5.0.

Destaca-se também que o método do cLHS balanceado foi o que apresentou menor variabilidade nos valores do índice Kappa obtidos pelos três classificadores avaliados, não diferindo estatísticamente entre si pelo Teste Z a 5% de probabilidade.

Analisando-se as matrizes de erro apresentadas nas Tabelas 4, 5 e 6, observa-se uma tendência geral que foi a concentração dos acertos na classe dominante, e os baixos valore de acerto para as classes menores GXbd, CXbd e RYbd.

Na Tabela 4 encontram-se as predições feitas pelo classificador Random Forest, é evidente a concentração das amostras de validação (200) na classe LVAd, que representa a classe dominante em termos de área. As classes de menor área (GXbd e RYbd) foram as piores predições, sendo que o classificador não acertou nenhuma amostra, salvo uma exceção no método cLHS. A classe CXbd também teve sua predição prejudicada, apresentando também só um acerto no método cLHS balanceado.

Tabela 4. Matrizes de confusão dos três métodos: classificador Random Forest

Random Forest – Amostragem Aleatória					
	CXbd	**GXbd**	**LVAd**	**PVAd**	**RYbd**
CXbd	0	0	2	0	0
GXbd	0	0	3	0	0
LVAd	17	7	146	11	0
PVAd	0	0	3	10	1
RYbd	0	0	0	0	0
Random Forest – Amostragem cLHS					
	CXbd	**GXbd**	**LVAd**	**PVAd**	**RYbd**
CXbd	0	0	1	0	0
GXbd	0	1	0	0	0
LVAd	17	6	142	23	1
PVAd	0	0	1	5	0
RYbd	0	0	0	0	0
Random Forest – Amostragem cLHS balanceado					
	CXbd	**GXbd**	**LVAd**	**PVAd**	**RYbd**
CXbd	1	0	0	0	0
GXbd	0	0	0	0	0
LVAd	9	6	147	5	0
PVAd	0	2	11	14	2
RYbd	0	0	0	0	0

A classificação feita com o GBM (Tabela 5) também apresenta um comportamento similar ao RF e C5.0, com nítida concentração de amostras na classe dominante (LVAd). O método cLHS balanceado foi penalizado principalmente por não ter acertado nenhuma amostra nas classes RYbd e GXbd. A classe PVAd foi melhor representada pelo método do cLHS balanceado, que apresentou o maior número de acertos nesta classe nos três classificadores.

Tabela 5. Matrizes de confusão dos três métodos: classificador GBM

GBM – Amostragem Aleatória					
	CXbd	**GXbd**	**LVAd**	**PVAd**	**RYbd**
CXbd	1	0	3	0	0
GXbd	0	1	2	1	0
LVAd	16	6	145	10	0
PVAd	0	0	4	10	1
RYbd	0	0	0	0	0
GBM – Amostragem cLHS					
	CXbd	**GXbd**	**LVAd**	**PVAd**	**RYbd**
CXbd	2	0	3	0	0
GXbd	0	2	1	0	0
LVAd	15	5	139	20	1
PVAd	0	0	1	8	0
RYbd	0	0	0	0	0
GBM – Amostragem cLHS balanceado					
	CXbd	**GXbd**	**LVAd**	**PVAd**	**RYbd**
CXbd	1	0	2	0	0
GXbd	0	0	0	0	0
LVAd	8	7	150	6	0
PVAd	1	1	6	13	2
RYbd	0	0	0	0	0

Tabela 6. Matrizes de confusão dos três métodos: classificador C5.0

C5.0 – Amostragem Aleatória					
	CXbd	**GXbd**	**LVAd**	**PVAd**	**RYbd**
CXbd	1	0	3	0	0
GXbd	0	1	2	0	1
LVAd	16	6	145	7	0
PVAd	0	0	3	13	1
RYbd	0	0	1	1	0
C5.0 – Amostragem cLHS					
	CXbd	**GXbd**	**LVAd**	**PVAd**	**RYbd**
CXbd	2	1	1	0	0
GXbd	0	0	1	1	0
LVAd	15	6	138	14	1
PVAd	0	0	4	9	0
RYbd	0	0	0	4	0
C5.0 – Amostragem cLHS balanceado					
	CXbd	**GXbd**	**LVAd**	**PVAd**	**RYbd**
CXbd	1	0	4	0	0
GXbd	0	1	0	1	0
LVAd	9	7	141	4	0
PVAd	0	0	13	14	2
RYbd	0	0	0	0	0

As Figura 6, 7 e 8 apresentam os mapas preditos pelos três classificadores para cada um dos métodos de amostragem Aleatório, cLHS e cLHS balanceado respectivamente. Em todos os nove mapas preditos só é possível identificar visualmente as classes LVAd, PVAd e CXbd, sendo que esta última não é identificável em todos os mapas. As classes GXbd e RYbd não são visualizáveis em nenhum dos mapas.

Nos mapas gerados pelo classificador RF (Figura 6), aquele obtido pelo método de amostragem cLHS foi de menor qualidade, visto que praticamente visualmente só apresenta uma classe, no caso a dominante LVAd. Já o mapa obtido com o cLHS balanceado evidencia de forma bem mais contínua a classe PVAd e também a classe CXbd, que se apresenta em pequenos polígonos espalhados na área.

Os mapas gerados pelo classificador GBM (Figua 7) praticamente não apresentam diferença visual, sendo as três classes representadas de maneira semelhante. Já na Figura 8 que apresenta os mapas gerados pelo classificador C5.0 existe uma nítida diferença de qualidade entre o mapa gerado pelo cLHS balanceado e os outros dois (cLHS e aleatório). O primeiro apresenta as unidades de mapeamento bem individualizadas e contínuas enquanto os demais têm predomínio expressivo da classe dominante juntamente com a classe PVAd que aparece fragmentada.

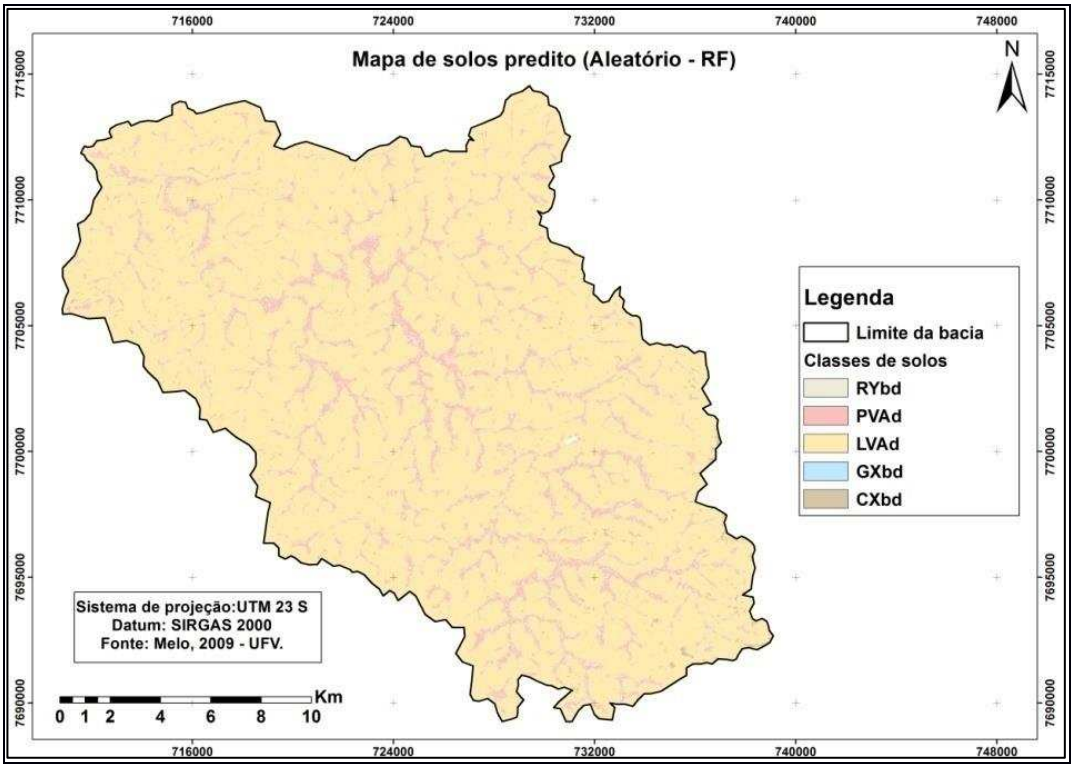


Figura 6. Mapa predito pelo método de amostragem aleatório com o classificador RF.

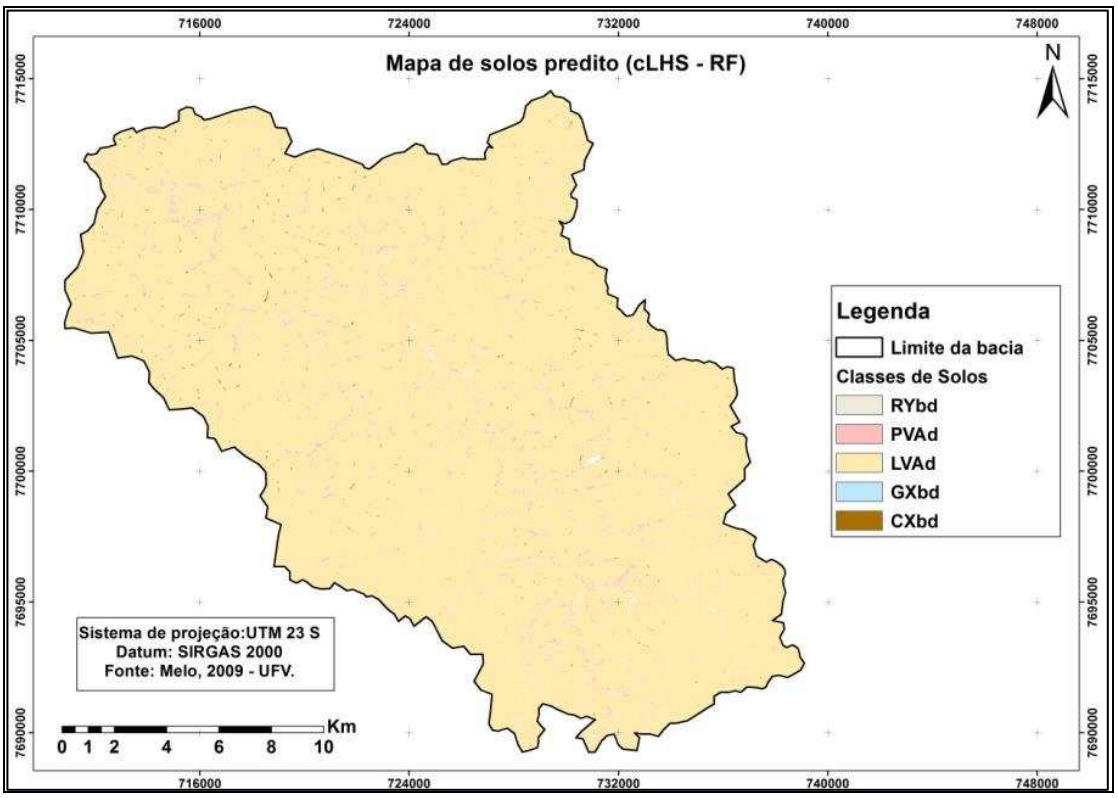


Figura 7. Mapa predito pelo método de amostragem cLHS com o classificador GBM.

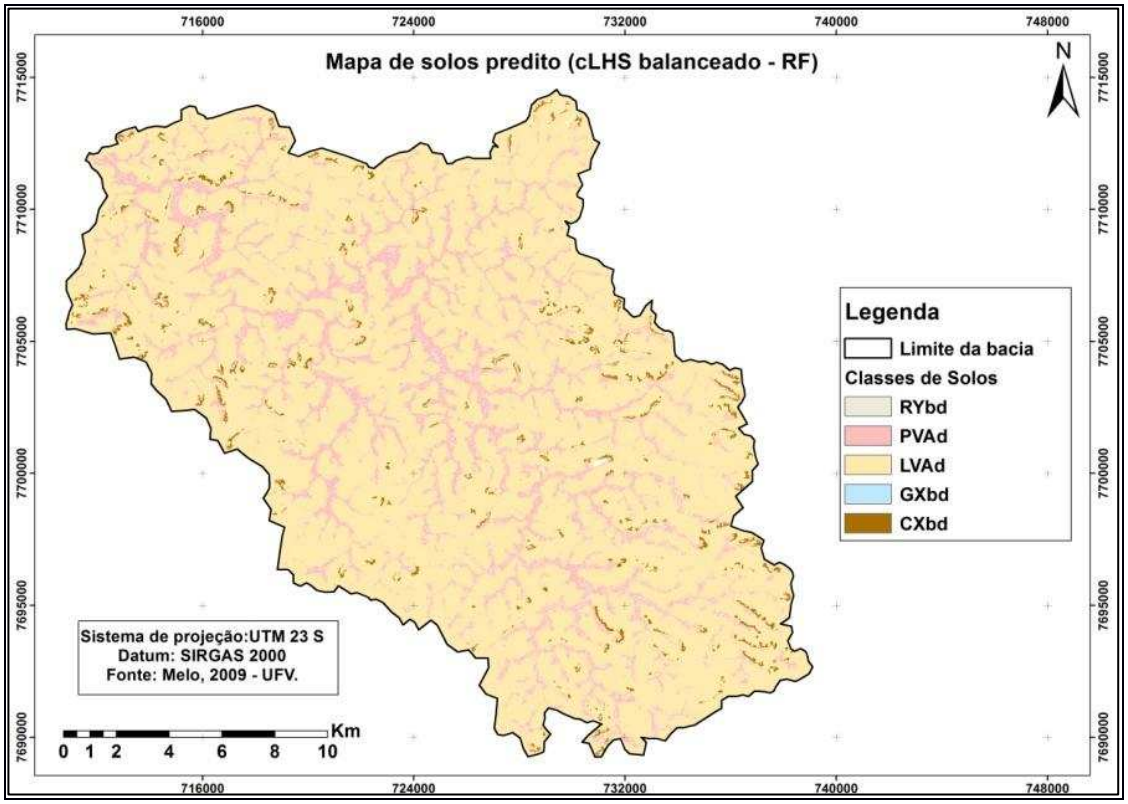


Figura 8. Mapa predito pelo método de amostragem cLHS balanceado com o classificador RF.

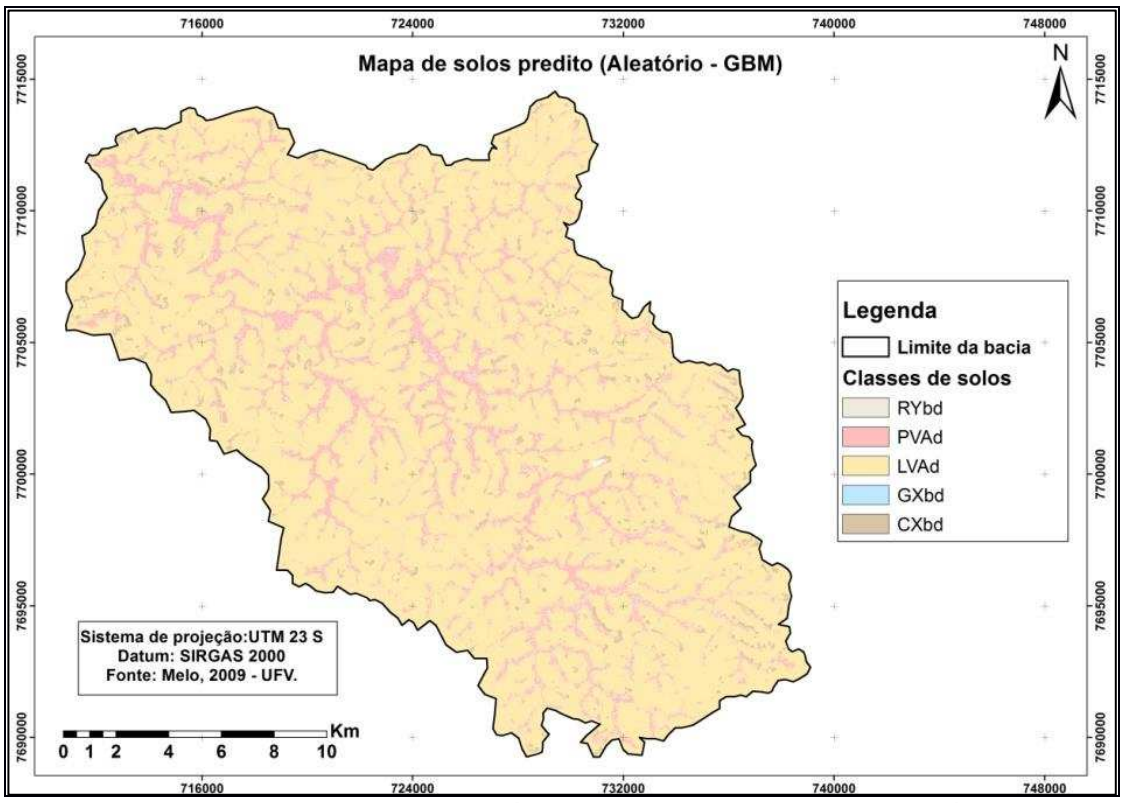


Figura 9. Mapa predito pelo método de amostragem aleatório com o classificador GBM.

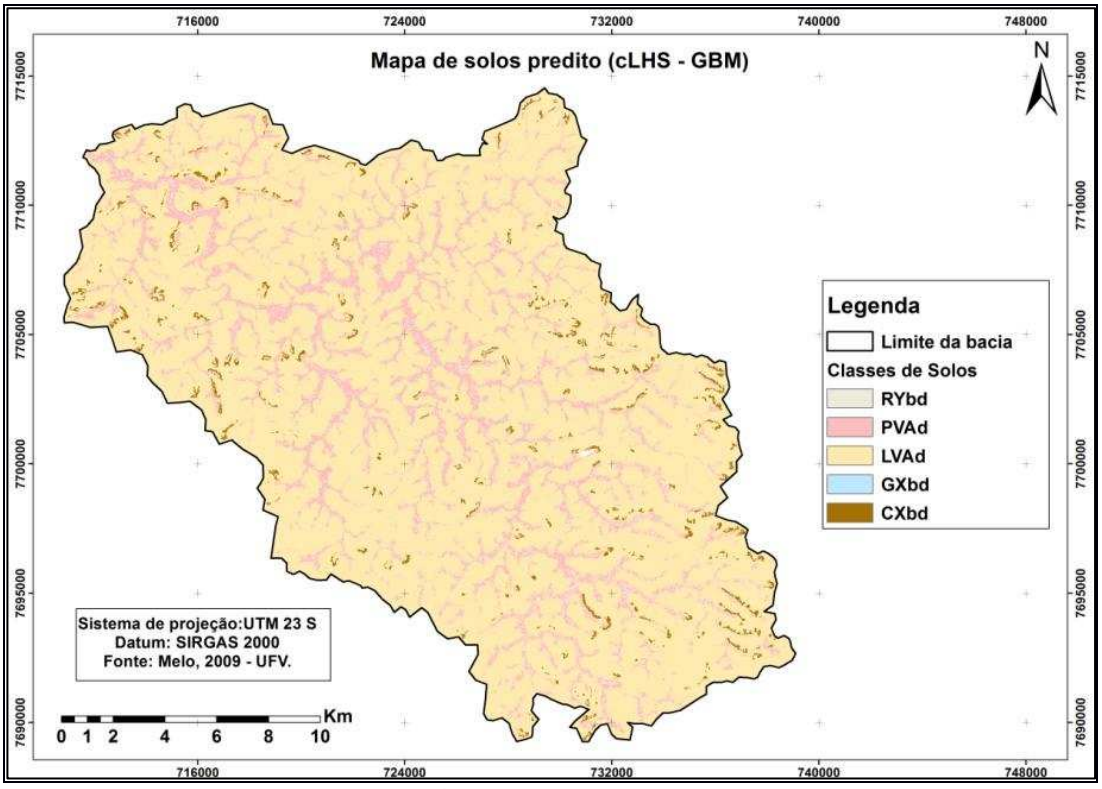


Figura 10. Mapa predito pelo método de amostragem cLHS com o classificador GBM.

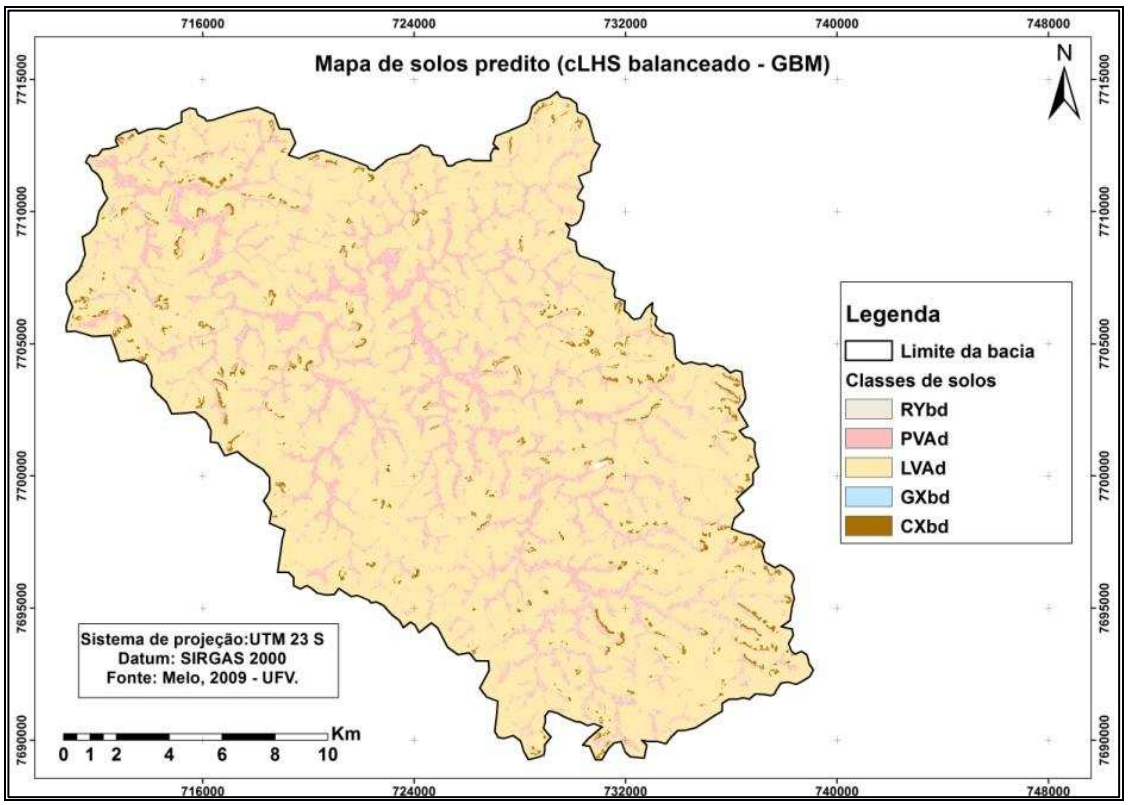


Figura 11. Mapa predito pelo método de amostragem cLHS balanceado com o classificador GBM.

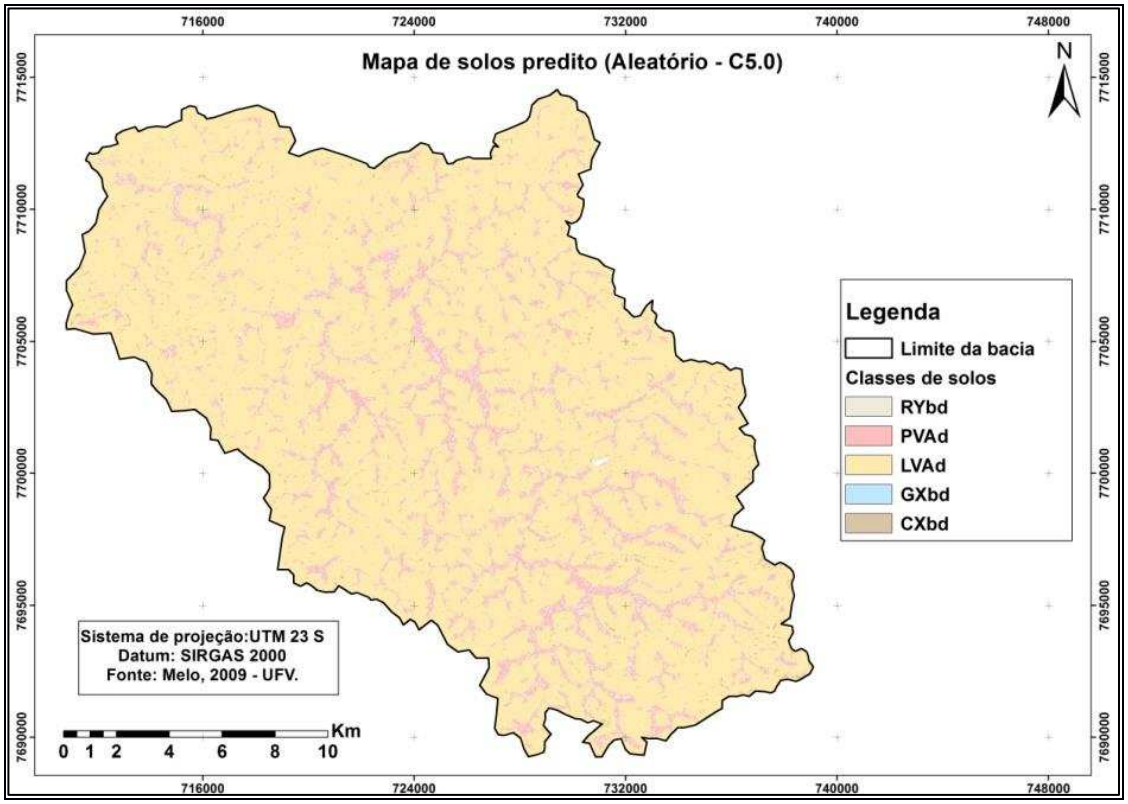


Figura 12. Mapa predito pelo método de amostragem aleatório com o classificador C5.0.

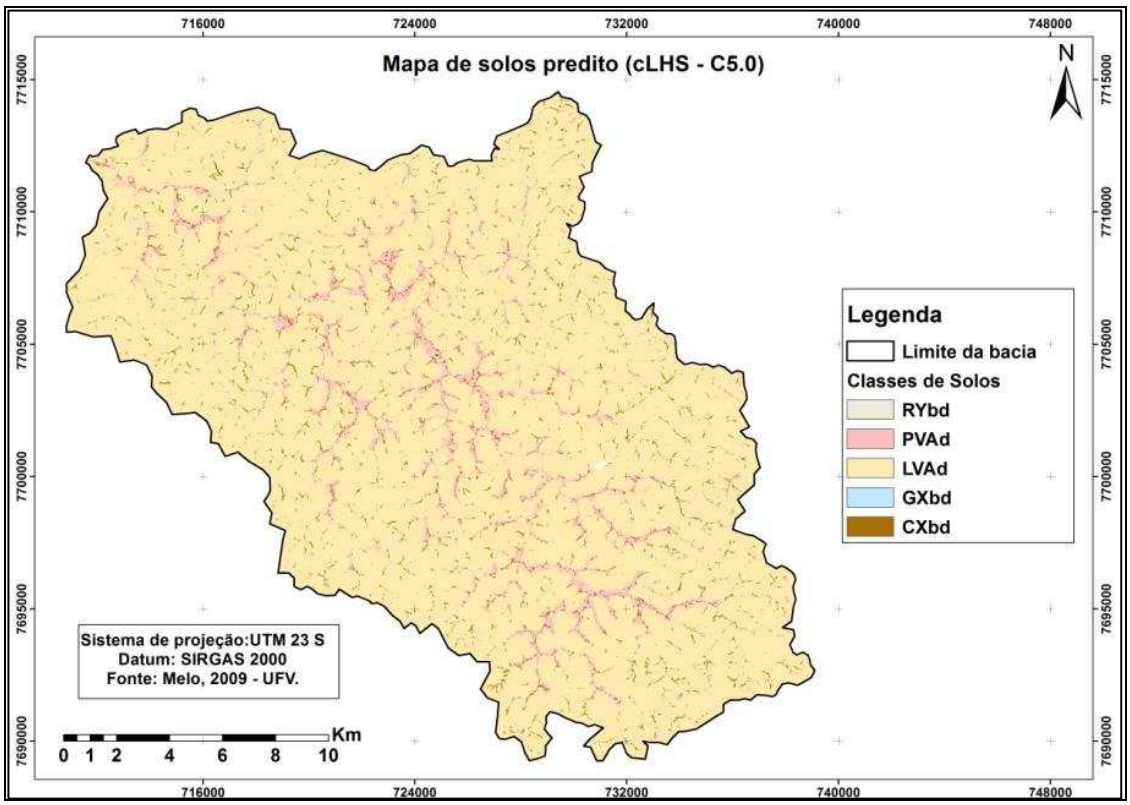


Figura 13. Mapa predito pelo método de amostragem cLHS com o classificador C5.0.

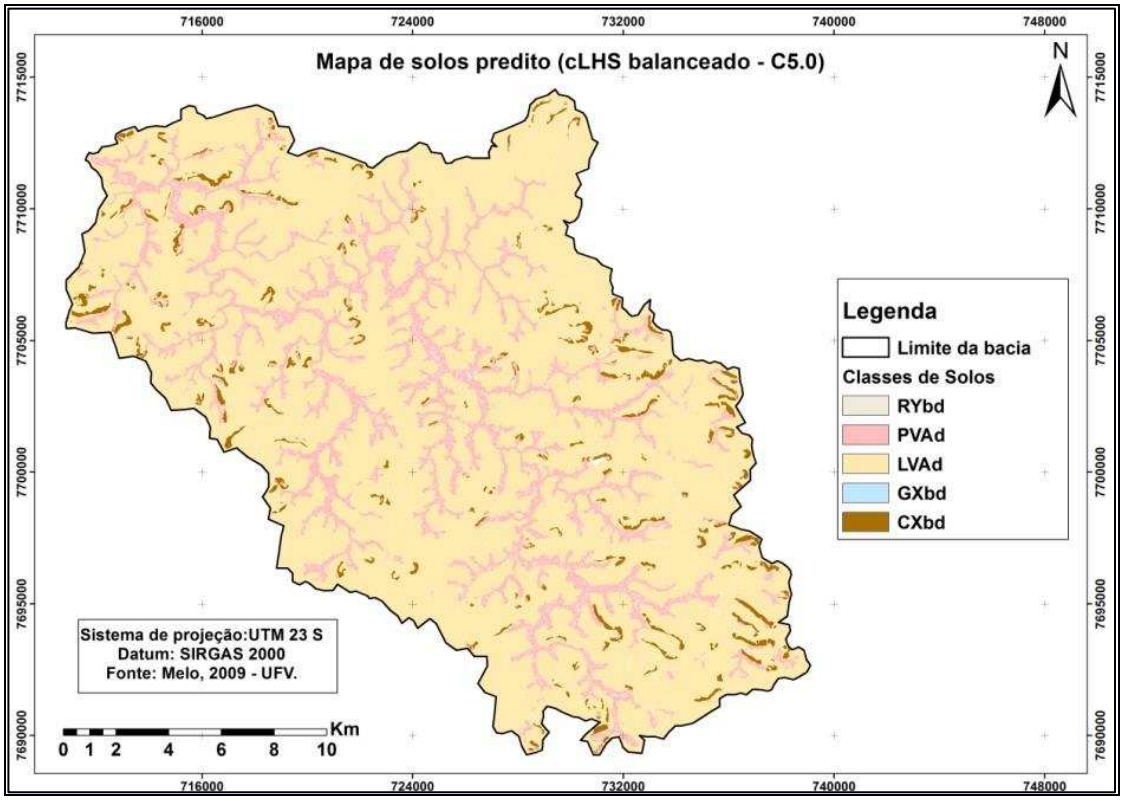


Figura 14. Mapa predito pelo método de amostragem cLHS balanceado com o classificador C5.0.

4. DISCUSSÃO

4.1. Distribuição das amostras nas classes preditas

Com relação à distribuição das amostras nas classes preditas pode-se considerar o método cLHS balanceado como o mais eficiente, visto que aumentou significativamente o número de amostras alocadas principalmente nas classes CXbd e PVAd. A baixa densidade de amostras na classe GXbd e a ausência das mesmas na classe RYbd foi atribuída ao fato de o balanceamento amostral ter sido feito a partir do mapa de solos legado. Neste estudo foram coletadas apenas três amostras para os Gleissolos em cada uma das malhas, já no estudo conduzido por Melo (2009), por exemplo, que trabalhou com mapeamento de solos por redes neurais também na bacia do rio Turvo Sujo, foram coletadas 15 amostras para esta mesma classe, possibilitando assim melhores condições de predição desta classe. Ainda referente à dificuldade encontrada para predizer a classe GXbd, relaciona-se o fato de que a maioria das amostras averiguadas em campo na porção mais baixa da paisagem, não eram efetivamente Gleissolos e sim Argissolos, sendo estes últimos membros de associação na unidade de mapeamento GXbd presente no mapa de solos legado. Da mesma forma os Neossolos Flúvicos não foram contemplados pelo método cLHS balanceado por que não existia uma unidade de mapeamento no mapa de solos legado onde estes solos eram o componente principal. Portanto o balanceamento poderia ter sido mais eficiente, se tivesse levado em consideração a ocorrência desta classe, presente somente como membro de associação na classe GXbd.

Isto proporciona uma reflexão sobre a importância de se balancear a amostragem, no entanto, no caso de se adotar a proposta aqui apresentada, é mito relevante compreender que a qualidade dos dados legados influirá diretamente na eficiência do balanceamento.

4.2. Diferenças entre os classificadores

A partir da utilização dos três classificadores aqui testados (RF, GBM e C5.0) não foi possível identificar o que tenha apresentado o melhor resultado

para todos os três métodos de amostragem. Este fato também foi observado por Brungard (2014) que após ter testado 20 diferentes classificadores, em três áreas de estudos diferentes, não conseguiu encontrar um que tenha tido o melhor desempenho nestas áreas.

O GBM e o RF apresentaram os maiores valores (0,42 e 0,41 respectivamente) para método cLHS balanceado não diferindo estatísticamente entre si pelo Teste Z a 5% de probabilidade. Já o C5.0 e o GBM apresentaram os melhores valores para os outros dois métodos (cLHS e aleatório). O pior resultado do índice Kappa obtido (0,15), foi pelo classificador RF para o método de amostragem cLHS. Este aspecto parece estar relacionado com o fato de que o método cLHS errou bastante nas classes CXbd e PVAd (Tabela 4), além das duas classes de menor área GXbd e RYbd, que não apareceram visualmente em nenhum dos mapas preditos. Este fato leva a crer que o desempenho pouco satisfatório, em termos de índice Kappa, está associado principalmente ao erro destas duas classes.

4.3. Diferenças entre os métodos de amostragem

O comportamentos do método cLHS balanceado diferiu dos demais no tocante à distribuição de amostras nas classes a serem preditas, bem como na menor oscilação dos valores do índice Kappa obtidos pelos três classificadores. Estes fatos evidenciam o potencial do método, que apresentou os melhores valores de Kappa (0,41, 0,42 e 0,38) para os classificadores RF, GBM e C5.0 respectivamente. Estes valores estão condizentes com a média encontrada em estudos nacionais (0,48) segundo ten Caten et al. (2012). Está concordante também com os valores encontrados por Odgers et al. (2014), em sua proposta de desagregação harmoniosa de unidades de mapeaento. Estes valores de índice Kappa (~0,4) podem ser considerados como regulares conforme a classificação proposta por Landis E Kock (1977), que considera valores de 0,4 a 0,6 como regulares (moderate), 0,6 a 0,8 bons (substantial) e maiores que 0,8 como ótimos (almost perfect).

A cerca destes valores regulares, acredita-se que o principal fator responsável pelos mesmos, seja a dificuldade encontrada na classificação das classes de menor área GXbd, CXbd e principalmente RYbd, sendo que esta última só estava presente no mapa legado na forma de componente de

associação na unidade de mapeamento GXbd. Deste modo a mesma não foi contemplada na proposta de balanceamento apresentada neste estudo, e sendo assim representou uma dificuldade eminente de classificação pelos modelos preditivos.

Melo (2009) que trabalhou na bacia do rio Turvo Sujo mapeamendo solos com o uso de redes neurais, também encontrou dificuldades na detecção da classe dos Gleissolos, no entanto teve a classe dos Neossolos Flúvicos como uma das melhores classificadas. Este aspecto pode estar relacionado com a quantidade de amostras utilizadas para a classficação desta classe (RYbd). No caso do estudo desenvolvido por Melo (2009) foram coletadas 4 amostras de RYbd, sendo 1 perfil completo e 3 pontos de observação. Já no estudo aqui apresentado foram coletadas 2 amostras na malha aleatória, apenas 1 na malha cLHS e nenhuma no cLHS balanceada. Portanto embora seja uma diferença numericamente pequena, tudo indica que para estas classes de menor expressividade poucas amostras a mais podem ser decisivas no processo da classificação. Desta forma, isto pode ser um indicativo de que quando estas classes forem mal classificadas um pequeno incremento amostral pode trazer um ganho significativo no resultdo final.

Uma classe que possa talvez ser utilizada como um indicadora da eficiência dos métodos de amostragem, é a classes dos Cambissolos (CXbd). Isto porque é uma classe intermediária, não sendo tão pouco expressiva como GXbd e RYbd, nem tão expressiva quanto LVAd e PVAd. Portanto ao se analisar a presença da mesma nos mapas pode se constatar o quanto os métodos de amostragem foram eficientes para a classificação. Esta classe aparece expressivamente em todos os três mapas gerado pelo método do cLHS balanceado, e só está bem representada nos outros métodos pela classificação feita pelo GBM.

O método Aleatório apresentou melhores resultados do que o método cLHS em todas as classificações, este fato levanta a discussão sobre a eficiência generalizada do método de amostragem cLHS em distintas condições, visto que alguns trabalhos tem empregado este método de amostragem e encontrando bons resultados (MINASNY E MCBRATNEY, 2007; CARVALHO JUNIOR et al., 2014; BRUNGARD et al., 2010). No entanto como o método cLHS destaca-se pela eficiência em reprsentar a variabilidade existente entre as covariáveis ambientais (MINASNY E MCBRATNEY, 2006), e

como neste trabalho utilizou-se somente variáveis representativas do relevo, não obtendo-se outras que contemplassem outros fatores de formação de solos, pode ser que o método cLHS tenha tido seu efeito subestimado.

As matrizes de confusão evidenciam o quão distante a classificação ficou do informação de referência. É interessante ressaltar que praticamente todo o erro atribuído a classe CXbd, quer dizer todas as amostras que não foram classificadas como CXbd, foram classificadas como LVAd. Este aspecto está relacionado com a dificuldade de separar estas duas classes na área de estudo, já evidenciado no mapa de solos legado que possui sua maior unidade de mapeamento (LVAd2), composta por estas duas classes em associação. Obviamente um limitante a esta separação é a escala de trabalho aqui adotada (1:100.00), no entanto é necessário que se avance em metodologias ou no uso de covariáveis que auxiliem nesta separação, visando com isto que o MDS venha a de fato melhorar a informação de solos já existente.

O predomino da classe dos Latossolos, seguida pela dos Argissolos, encontrado neste trabalho, corrobora com os resultados encontrados por Melo (2009) que também encontrou predomínio de ambas em seu mapeamento por redes neurais artificiais. Entretanto uma diferença pronunciada entre estes dois trabalhos encontra-se na detecção dos Neossolos Flúvicos, que praticamente não foram identificados neste trabalho, e já no realizado por Melo (2009) representaram cerca de 5% da bacia. Neste contexto destaca-se que o resultado obtido com o mapeamento feito com redes neurais, foi mais eficiente, visto que realmente estes solos ocorrem com certa expressividade na área de estudo.

A partir dos resultados expressos nas matrizes de confusão é possível ainda perceber que embora o método cLHS balanceado não tenha estabelecido amostras nesta classe (RYbd), em todos os três classificadores as duas únicas amostras que existiam foram confundidas com PVAd. Este aspecto demonstra que embora a classe não tenha sido predita, foi confundida com uma classe próxima a ela. Enquanto nos resultados obtidos com a amostragem cLHS pelos três classificadores a classes RYbd foi confundida com LVAd.

Este tipo de análise de matrizes de erros permite uma reflexão interessante sobre a proporção dos erros cometidos pelos classificadores. Trata-se de avaliar a distância taxonômica ou ambiental existente entre as

classes. No caso de uma amostra que deveria ser classificada como RYbd e é classificada como LVAd, o erro deveria ser mais severamente penalizado, visto que são classes muito distantes do ponto de vista taxonômico e neste caso também ambiental. Isto porque os Latossolos e os Neossolos Flúvicos ocupam porções bem distintas da paisagem na área de estudo. Por outro lado o erro em uma amostra que deveria ser classificada como CXbd, e é classificada como LVAd, nesta área de estudo particularmente onde estas classes são ambientalmente próximas (CORRÊA, 1984), deveria ser um erro atenuado. Desta forma acredita-se ser interessante estabelecer e utilizar estes parâmetros, de distância taxonômica ou ambiental, como condicionadores dos erros, atenuando ou acentuando os mesmos conforme a semelhança ou discrepância existente entre as classes a serem preditas. Minasny e McBratney (2007) apresentam esta idéia em um trabalho de mapeamento de classes de solo na Austrália. Os autores elaboraram um quadro de distância taxonômica para as classes de solo do sistema de classificação australiano e associaram o mesmo no processo de predição através do uso de árvores de classificação.

5. CONCLUSÕES

- O método cLHS balanceado foi eficiente na distribuição das amostras dentre as classes a serem preditas, desfavorecendo a classes de maior área (LVAd) e favorecendo principalmente as classes CXbd e PVAd. No entanto não favoreceu significativamente as duas classes de menor área GXbd e RYbd.
- Embora a inserção do desbalanceamento das classes como um custo no método de amostragem cLHS tenha apresentado diferença em relação aos métodos não balanceados, o fato de ter sido feita a partir das unidades de mapeamento do mapa de solos legado, que continham mais de uma classe em associação, reduziu a eficiência da amostragem deste método .
- Os melhores resultados de índice Kappa obtidos dentre todas as combinações estabelecidas foram com o método de amostragem cLHS balanceado pelos classificadores Gradient Boosting Machine (GBM) e Random Forest (RF).

– O método de amostragem que apresentou resultados com menor oscilação nos valores do índice Kappa foi o cLHS balanceado, enquanto o que apresentou maior oscilação foi o cLHS.

6. REFERÊNCIAS BIBLIOGRÁFICAS

BRUNGARD, C.W.Advancing Digital Soil Mapping and Assessment in Arid Landscapes. Utah State University, All Graduate Theses and Dissertations. Paper 3305. 2014.

BRUNGARD, C.; BOETTINGER, J.L. Conditioned Latin Hypercube Sampling: Optimal Sample Size for Digital Soil Mapping of Arid Rangelands in Utah, USA. **Media**, n. December, 2010.

CARRÉ, F.; MCBRATNEY, A. B.; MINASNY, B. Estimation and potential improvement of the quality of legacy soil samples for digital soil mapping. **Geoderma**, v.141, n.1-2, p.1–14, 2007.

CHAWLA, N.V; JAPKOWICZ, N.; DRIVE, P. Editorial : Special Issue on Learning from Imbalanced Data Sets. **ACM SIGKDD Explorations Newsletter**, v.6, n.1, p.1-6, 2004.

CODEMIG - COMPANHIA DESENVOLVIMENTO ECONÔMICO DE MINAS GERAIS– MAPA GEOLÓGICO (www.portaldageologia.com.br) acesso em 26/05/2016. 2014.

CONGALTON, R.G.; GREEN, K. **Assessing the accuracy of remotely sensed data: principles and practices**. New York: Lewis Publishers, 1999, 160p.

CPRM–COMPANHIA DE PESQUISA DE RECURSOS MINERAIS - Mapa Geodversidade (www.cprm.gov.br/geobank) acesso em 26/05/2016. 2016.

CORRÊA, G.F. Modelo de evolução e mineralogia da fração argila de solos do planalto de Viçosa, MG. Dissertação. (Mestrado em Solos e Nutrição de Plantas) – Universidade Federal de Viçosa. Viçosa-MG, 1984, 86.

KEMPEN, B. et al. Updating the 1:50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. **Geoderma**, v.151, n.3-4, p.311–326, 2009.

LAGACHERIE, P., MCBRATNEY, A.B., Chapter 1. Spatial soil information systems and spatial soil inference systems: perspectives for Digital Soil Mapping. In: P. Lagacherie, A.B. McBratney and M. Voltz (Editors), **Digital Soil Mapping: An Introductory Perspective**. 2007.

LANDIS, J.R.; KOCK, G.G. The mesurament of observer agreement for categorical data. Biometrics. 1977, 33, 159-174.

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

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

MOSLEH, Z. et al. The effectiveness of digital soil mapping to predict soil properties over low-relief areas. **Environmental Monitoring and Assessment**, v.188, n.3, p.1–13, 2016.

MELO, L.V. Uso de redes neurais artificiais no mapeamento de solos na Bacia do Rio Turvo Sujo – Viçosa MG. Viçosa, 2009, 80p.

NKWUNONWO, U. C. Digital Mapping of Soil Chemical Properties for Erosion Hazard Management in Bayelsa State , Nigeria. v.4, n.13, p.209–217, 2015.

ODGERS, N.P.; SUN, W.; McBRATNEY, A.B.; MINASNY, B.; CLIFFORD, D. Disaggregating and harmonising soil map units through resampled classification trees. **Geoderma.** 214 (2014) 91–100.

PAHLAVAN-RAD, M. R. et al. Legacy soil maps as a covariate in digital soil mapping: A case study from Northern Iran. **Geoderma**, v.279, p.141–148, 2016.

PÁSZTOR, L.; SZABÓ, J.; BAKACSI, Z. Digital processing and upgrading of legacy data collected during the 1:25 000 scale Kreybig soil survey. **Acta Geodaetica et Geophysica Hungarica**, v.45, n.1, p.127–136, 2010.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. 2016.

ROUDIER, P.; HEWITT, A.E. & BEAUDETTE, D.E. A conditioned latin hypercube sampling algorithm incorporating operational constraints. In: MINASNY, B.; MALONE, B.P. & McBRATNEY, A.B., eds. Digital soil assessments and beyond. London, CRC Press/Balkema, 2012. p.227-232.

TEN CATEN, A. et al. Mapeamento digital de classes de solos : características da abordagem brasileira. **Ciência Rural**, v. 42, p. 1989–1997, 2012.

TESKE, R.; GIASSON, E.; BAGATINI, T. Comparação De Esquemas De Amostragem Para Treinamento De Modelos Preditores No Mapeamento Digital De Classes De Solos. **Revista Brasileira de Ciência do Solo**, v.39, n.1, p.14–20, 2015.

VAYSSE, K.; LAGACHERIE, P. Evaluating Digital Soil Mapping approaches for mapping GlobalSoilMap soil properties from legacy data in Languedoc-Roussillon (France). **Geoderma Regional**, v.4, p.20–30, 2015.

ZHANG, S.-J. et al. An heuristic uncertainty directed field sampling design for digital soil mapping. **Geoderma**, v.267, p.123-136, 2016.

CAPITULO 3

SELEÇÃO DE COVARIÁVEIS PARA PREDIÇÃO DE UNIDADES DE MAPEAMENTO DE SOLOS NO QUADRILÁTERO FERRÍFERO, MINAS GERAIS

RESUMO

VASCONCELOS, Bruno Nery Fernandes, D.Sc., Universidade Federal de Viçosa, setembro de 2016. **Seleção de covariáveis para predição de unidades de mapeamento de solos no quadriátro ferrífero, Minas Gerais.** Orientador: Elpídio Inácio Fernandes Filho. Coorientadores: João Carlos Ker e Carlos Ernesto G.R. Schaefer.

Os mapas de solo são ferramentas essenciais de planejamento e uso da terra, e a demanda por informações cada vez mais detalhadas nestes mapas cresce concomitantemente com a disponibilidade e aprimoramento das bases de dados existentes. Atualmente acredita-se que a produção de mapas mais detalhados seja potencializada com a aplicação de geotecnologias associadas aos trabalhos de campo. O termo utilizado para a produção destes mapas é mapeamento digital de solos (MDS), que visa estabelecer a distribuição espacial dos solos em uma determinada área através de modelos matemáticos. As propostas de trabalho com MDS se baseiam em combinações entre algorítimos matemáticos, métodos de validação e diversos conjuntos de covariáveis preditoras. Os principais objetivos deste trabalho foram avaliar os métodos de seleção de variáveis: Percentil, Correlação, Índice Ginni e diminuição média da exatidão (Mean decrease Accuracy-MDA), avaliar o potencial do uso de dados geofísicos como covariáveis ambientias, e identificar o menor conjunto de covariáveis que forneça melhores resultados na predição de classes de solo na região do Quadrilátero Ferrífero, Estado de Minas Gerais. Para tanto foi utilizado um conjunto de 141 variáveis associadas aos fatores de formação dos solos relevo, clima, organismos e material de origem, derivadas respectivamente do SRTM (90 m), dados climáticos do world clim, Landsat 8, mapa geológico e dados geofísicos. A partir dos conjuntos de variáveis selecionados foi feita a predição dos solos, em quatro níveis de detalhamento

(9, 24, 34 e 75 Unidades de Mapeamento - UMs), posteriormente os mapas foram comparados pixel a pixel com o mapa de solos de referência. Os resultados mostram que o processo de seleção de variáveis foi eficiente em manter a exatidão dos mapas com aproximadamente 10% das variáveis iniciais. O mapa predito pelo Random Forest que mais se aproximou do mapa de referência foi elaborado pelas covariáveis selecionadas pelo Percentil para 34 UMs de solo. A variabilidade no conjunto de variáveis selecionada fortalece a importância de se utilizar variáveis vinculadas a mais de um fator de formação de solos. As covariáveis geofísicas de gamaespectometria foram selecionadas com frequência por diferentes métodos, evidenciando o potencial ainda pouco explorado que possuem como covariáveis preditoras em MDS.

Palavras-chave: Random Forest, MDS, índice kappa, Percentil, Índice Ginni.

ABSTRACT

VASCONCELOS, Bruno Nery Fernandes, D.Sc., Universidade Federal de Viçosa, September, 2016. **Selection of covariates for prediction of soil mapping units in the Quadrilátero Ferrífero, Minas Gerais: methodological approach.**Adviser: Elpídio Inácio Fernandes Filho. Co-advisers: João Carlos Ker and Carlos Ernesto G. R. Schaefer.

Soil maps are essential tools for planning of land use, and the demand for increasingly detailed information on these maps grows concurrently with the availability and improvement of existing databases. Currently it is believed that the production of more detailed maps is only possible with the use of geotechnologys associated with the field work. Currently the term used for the production of these maps is Digital Soil Mapping (DSM), which aims to establish the spatial distribution of soil in a particular area using mathematical models. The work proposals with DSM are based on combinations of mathematical algorithms, validation methods and different sets of predictor variables. The main objectives of this study are to evaluate the variable selection methods: Percentil, Correlation, Ginni index and Mean Decrease Accuracy (Mean decrease Accuracy-MDA), and identify the set with fewer variables to provide better results in the class prediction soil in the area of the Quadrilátero Ferrífero,

State of Minas Gerais. Therefore a set of 141 variables associated with the soil formation factors: relief, climate, organisms and parent material, derived respectively from SRTM (90 m), climatic data of worldclim, Landsat 8, geological map and geophysical data was submitted the four methods of selecting variables. From the sets of selected variables was the prediction of soil made in four levels of detail (9, 24, 34 and 75 Mapping Units - MUs), then the maps were compared pixel by pixel with the reference soil map (Legacy data). The results show that the variable selection process was effective in maintaining the accuracy of maps with approximately 10% of the initial variables. The map closest to the reference map was prepared by Percentil method for 34 MUs soil. The variability in the selected set of variables strengthens the importance of using variable is bound to more than one soil formation factor. The geophysical variables gama ray spectometry were selected with high frequency by different methods, highlighting the potential still little explored that have as predictor variables in MDS.

Keywords: Random Forest, MDS, índice kappa, Percentil, Índice Ginni, Mean Decrease Accuracy.

1. INTRODUÇÃO

Atualmente a demanda por mapas mais detalhados de classes ou atributos de solos, que possam ser utilizados para fins de manejo e conservação dos mesmos é cada vez mais urgente (DALMOLIN e TEN CATEN, 2015), sendo que a produção destes dados pode ser obtida por uma interação entre os conhecimentos tradicionais de levantamento de solos, e as novas tecnologias computacionais, disponíveis atualmente, capazes de trabalhar com bases de dados cada vez mais detalhadas.

A necessidade de tornar os mapeamentos de solo cada vez mais quantitativos, visando obter informações numéricas sobre a distribuição dos mesmos na paisagem, gerou uma linha de pesquisa conhecida como Mapeamento Digital de Solos (MDS). A predição de classes ou propriedades do solo através do MDS tiveram sua concepção metodológica no trabalho apresentado por McBratney et al., (2003), e posteriormente tiveram sua definição aprimorada como “a criação de sistemas de informação espacial de

solos por meio de modelos numéricos, visando a inferir as variações espaciais e temporais de classes e propriedades do solo, a partir de observações, conhecimento e dados de covariáveis ambientais relacionadas" (LAGACHERIE e McBRATNEY, 2007).

Os estudos realizados e publicados sobre MDS vem crescendo rapidamente na comunidade acadêmica e diante deste contexto ten Caten et al. (2012) elaboraram uma extensa revisão bibliográfica sobre os trabalhos desenvolvidos no Brasil, no período de 2006 a 2012. Os autores reuniram as principais características dos trabalhos já realizados, visando gerar perspectivas mais amplas sobre o caminho que o Mapeamento Digital de Solos vem tomando no país, além de identificar lacunas metodológicas ainda existente nos métodos. Um dos principais pontos levantados por ten Caten et al. (2012) é a discussão sobre as covariáveis utilizadas para alimentar os modelos preditivos. As variáveis estão associadas aos fatores de formação de solos definidos desde os estudos de Dokuchaev, em meados do século XIX, equacionados posteriormente por Jenny (1941) (S = f (ClORPT)), onde (Cl) é clima, (O) organismos, (R) relevo, (P) material de origem, e (T) tempo. Nesta revisão ten Caten et al. (2012) identificaram que as variáveis associadas ao relevo foram utilizadas na totalidade dos trabalhos avaliados. Já as variáveis associadas a clima, organismos e a material de origem foram menos presentes, sendo contempladas em cerca de 25% dos trabalhos avaliados. Essa diferença é atribuída à importância do fator relevo na formação dos solos, expressa na estreita relação existente entre as diversas variáveis derivadas do relevo e a distribuição dos solos. Mas também ao fato de que atualmente existem boas bases de dados, em diferentes resoluções espaciais, na forma de Modelos Digitais de Elevação (MDEs) disponíveis gratuitamente.

As variáveis associadas ao relevo permanecem sendo as mais utilizadas, com destaque para Declividade, Curvatura Plana, Curvatura de Perfil e Índice de Umidade Topográfica (TESK, 2015; VAYSSE e LAGACHERIE, 2015; GIASSON et al., 2015, DIAS, 2015; BAGATINNI, 2015; FONNESBECK, 2015; MOSLEH et al., 2106; TAGHIZADEH MERJARDHI, 2016). Outras variáveis de relevo que têm sido utilizada com menos frequência são Face de exposição (Aspect), MRVBF (Multi-resolution valley bottom flatness), MRRTF (Multi-resolution ridge top flatness), Stream Power Index (SPI), Flow Direction e Flow Accumulation.

Destaca-se que existe a necessidade de se explorar melhor o potencial de covariáveis ambientais vinculadas aos demais fatores de formação de solos, visto que estas têm sido ainda pouco utilizada. As covariáveis relacionadas ao material de origem por exemplo, foram encontradas em 4 dos 10 trabalhos consultados, sendo estas reduzidas basicamente a quatro variáveis, mapas geológicos, índice Clay minerals, Iron oxides e Salinity ratio é (VAYSSE e LAGACHERIE, 2015; DIAS, 2015; TAGHIZADEH-MERJARDHI, 2016). Quanto aos aspectos biológicos destaca-se o uso do Normalized Difference Vegetation Index (NDVI) e de mapas de uso e cobertura de solos (VAYSSE e LAGACHERIE, 2015; DIAS, 2015; TAGHIZADEH-MERJARDHI, 2016). As variáveis vinculadas ao clima foram empregadas no MDS somente em um trabalho dos avaliados com destaque para temperaturas máximas e mínimas, e precipitação pluviométrica (VAYSSE e LAGACHERIE, 2015).

Para Lagacherie (2008) uma perspectiva de pesquisa encontra-se na busca por novas covariáveis, que podem ser obtidas pelo processamento de covariáveis simples ou através de inputs de novas fontes de dados.

A aerogamaespectometria que une métodos geofísicos e técnicas de Sensoriamento Remoto, informa da radiação gama emitida por rochas e solos, sendo comumente utilizada para mapeamentos geológicos em área densamente vegetadas. Os raios gama tem comprimentos de onda em torno de 10^{-3}nm e são originados pelo decaimento dos isótopos radioativos K^{40}, Tl^{208} (Th) e Bi^{214} (U) contidos em minerais (WILFORD e MINTY, 2007).

A presença destes elementos radioativos nas rochas é diferente da presença dos mesmos nos solos. Nas rochas o K está presente em maiores quantidades em rochas ígneas ácidas como granito, rhyolito e pegmatito. Por outro lado, é praticamente ausente em rochas mais básicas como basaltos, serpentinitos e peridotitos. O urânio (U) ocorre de duas formas principais, U^{+4} e U^{+6}. A forma oxidada U^{+6}forma complexos com oxigênio ($UO2^{+2}$). Estes são íons móveis que tipicamente formam complexos químicos com carbonatos, sulfatos e cloretos. Já a forma mais reduzida U^{+4} é insolúvel e é componente de minerais primários.

O tório (Th) ocorre somente na valencia +4, sendo, portanto, não susceptível a mudanças em condições redox. A quantidade de tório solúvel é geralmente muito pequena, sendo que este se torna solúvel em condições de pH's muito baixos ou em pH neutro, quando está associado a complexos

orgânicos. Urânio e tório são comumente encontrados nos solos como acessórios em minerais resistentes como zircão, titanita, allanita e monazita (WILFORD e MINTY, 2007).

A radioatividade nos solos está diretamente ligada ao material de origem e aos processos pedogenéticos predominantes em sua formação, pois durante os processos de intemperismo os radioelementos são liberados das rochas e redistribuídos nos solos (WILFORD e MINTY, 1997). Em solos pouco desenvolvidos como Neossolos Litólicos ou Cambissolos, originários de rochas máficas, observam-se poucas modificações dos radioelementos, porém em solos mais intemperizados, provenientes destas mesmas rochas pode se encontrar concentrações de duas a três vezes maiores do que no material de origem (SANTOS et al., 2008).

Alguns estudos já empregaram os dados gamaespectométricos em monitoramentos ambientais e na pesquisa de solos. Bachi et al. (1999) utilizaram a técnica de atenuação de radiação gama para determinar umidade e densidade do solo. Vaz et al. (1999) utilizaram esta mesma técnica para determinar tamanho de partículas de solo. Rebelo (2000) observou que o elemento Th permanece fixo no solo durante o intemperismo e a pedogênese. Becegato e Ferreira (2005) utilizaram dados de gamaespectometria no entendimento da distribuição de fertilizantes fosfatados em cultura de soja, trigo e cana de açúcar no noroeste do Paraná.

Alguns trabalhos com gamaespectometria aplicada ao estudo de solos, observaram bons resultados de correlação entre as propriedades dos solos e estes dados. Bierwirth (1996) encontrou relações significativas entre as concentrações de K obtidas em dados de aerogamaespectometria e valores de pH, em diferentes substratos minerais. Em um trabalho realizado na porção ocidental da Austrália, Taylor et al. (2002) estudaram correlações entre propriedades físicas e mineralógicas dos solos com dados de gamaespectometria, encontraram resultados que acentuam o potencial de uso desta informação no estudo de solos. Os autores encontraram elevadas concentrações de Th e U em áreas de concreções lateríticas, associados a baixas concentrações de K. Os solos arenosos com predominância de quartzo na fração areia apresentaram baixos valores dos elementos radioativos. A presença de hematita foi correlacionada com altos valores de Th.

Herrmann et al. (2010), em um trabalho realizado no norte da Tailândia, avaliaram o potencial da espectometria de raios gama para auxiliar no mapeamento convencional de solos em uma região muito montanhosa e coberta por floresta tropical densa, dificultando assim os trabalhos de campo. Os resultados apontam que o método gamaespectométrico apresenta um enorme potencial para auxiliar mapeamentos de solos mais detalhados em regiões de relevo mais acidentado e com heterogeneidade geológica.

Diante da possibilidade de se utilizar outras covariáveis que apresentem potencial na predição de classes e ou propriedades de solos, torna-se necessário avaliar este potencial junto às demais covariáveis existentes para cada área de trabalho. Dessa forma a aplicação da seleção de variáveis é muito relevante em MDS, pois permite identificar entre um conjunto significativo de covariáveis possíveis de serem utilizadas nos modelos preditivos, aquele que melhor representa a condição dos solos a serem mapeados. O mapeamento convencional de solos produz mapas utilizando poucas fontes de informação, geralmente ligadas ao relevo e ao material de origem. Portanto na reconstrução do modelo mental de distribuição dos solos elaborado pelo pedólogo, deve-se buscar um conjunto robusto e simples de variáveis, visando tornar a abordagem do MDS mais próxima daquela até então empregada no mapeamento de solos convencional.

Em alguns estudos sobre dados espectrais vinculados a ciência do solo como os conduzidos por Jia et al. (2016) e Vohland et al. (2014), que utilizaram espectroscopia de infravermelho próximo para determinar algumas propriedades químicas e biológicas dos solos tais como pH, Nitrogênio, Carbono Orgânico e biomassa microbiana, o processo de seleção de covariáveis é amplamente aceito como um passo importante na predição.

O Random Forest (RF) é uma técnica de inteligência artificial baseada em arvóres de regressão utilizada em análises multivariadas. O RF fornece uma estimativa de erro, usando (out-of-bag) OOB, onde um terço dos dados de entrada é excluído da predição para serem usados na estimativa do erro de predição. A partir da diferença do erro encontrada no OOB o RF calcula a importância das variáveis e este tem sido um parâmetro utilizado para selecionar variáveis (BREIMAN, 2001). Menze et al. (2009) avaliaram a eficácia do ranqueamento de importância do RF através do Índice Ginni, na seleção de dados espectrais em comparação a outros dois métodos quimiométricos

Regressão de Componentes Principais (PCR) e “Partial least squares regression” (PLS). Os autores constataram que o Random Forest a partir do Índice Ginni apresentou um bom desempenho na seleção dos melhores parâmetros dentro do conjunto total de dados, sugerindo o uso do mesmo antes de se aplicar um dos outros dois métodos.

Utilizando também o ranqueamento produzido pelo RF, Brungar (2014) realizou um estudo de mapeamento digital de classes de solo em três regiões de clima árido nos Estados Unidos com um conjunto total de 113 variáveis preditoras. Após um processo de seleção de variáveis o autor chegou a um conjunto ideal de variáveis para cada área de trabalho, sendo oito o número máximo de variáveis selecionadas para cada uma das três áreas de trabalho.

Outros métodos de seleção de variáveis estão sendo testados em trabalhos de MDS, como no trabalho de Miller, et al. (2015) que utilizou o software de mineração de dados Cubist em um conjunto inicial de 412 variáveis representando diferentes fatores de formação de solos. Os autores selecionaram diferentes sub-conjuntos de 5 a 14 variáveis que produziram modelos com exatidão (R^2) variando entre 0,21 e 0,93. Pontes et al., (2009) trabalharam em uma proposta metodológica de classificação de amostras de solo a partir de Espectroscopia de desagregação induzida por laser (LIBS) associada a técnicas quimiométricas. Para tanto os autores avaliaram três técnicas, Algorítimo de Progressão Sucessiva (SPA), Algorítimo Genetico (GA) e Formulação Gradual (SW). Neste estudo os autores constaram que GA apresentou o pior desempenho, selecionando o maior número de variáveis (17), enquanto SW e SPA selecionaram sete e cinco variáveis respectivamente.

Neste contexto o principal objetivo deste trabalho foi avaliar duas novas formas de selecionar covariáveis: a primeira a partir da remoção de covariáveis com alta correlação não linear, e a segunda através do uso do Percentil para estabelecer as melhores covariáveis. E compará-las ao processo de seleção de covariáveis presente no Random Forest, a partir de seus dois Índices (Ginni e Mean Decrease Accuracy). Além disso, buscou-se avaliar a contribuição no potencial preditivo do modelo dos dados climáticos, (wordlclim) e dos dados aerogeofísicos de gamaespectormetria, visando obter o menor conjunto de variáveis capaz de predizer com maior exatidão, unidades de mapeamento de solos em diferentes níveis de detalhamento, na Área de Proteção Ambiental Sul (APA-Sul), localizada no Quadrilátero Ferrífero, Estado de Minas Gerais.

2. MATERIAL E MÉTODOS

2.1. Área de estudo

O estudo foi realizado para a Área de Proteção Ambiental Sul (APA-Sul) que se localiza na região centro sul do Estado de Minas Gerais (Figura 1), na província geológica do Quadrilátero Ferrífero. A criação da Área de Proteção Ambiental na região sul de Belo Horizonte, em 26 de julho de 2001, através da Lei Estadual no 13.960, fundamentou-se na peculiaridade dos atributos dos meios físico e biótico, além de aspectos estéticos, culturais e econômicos e o imenso potencial hídrico da região.

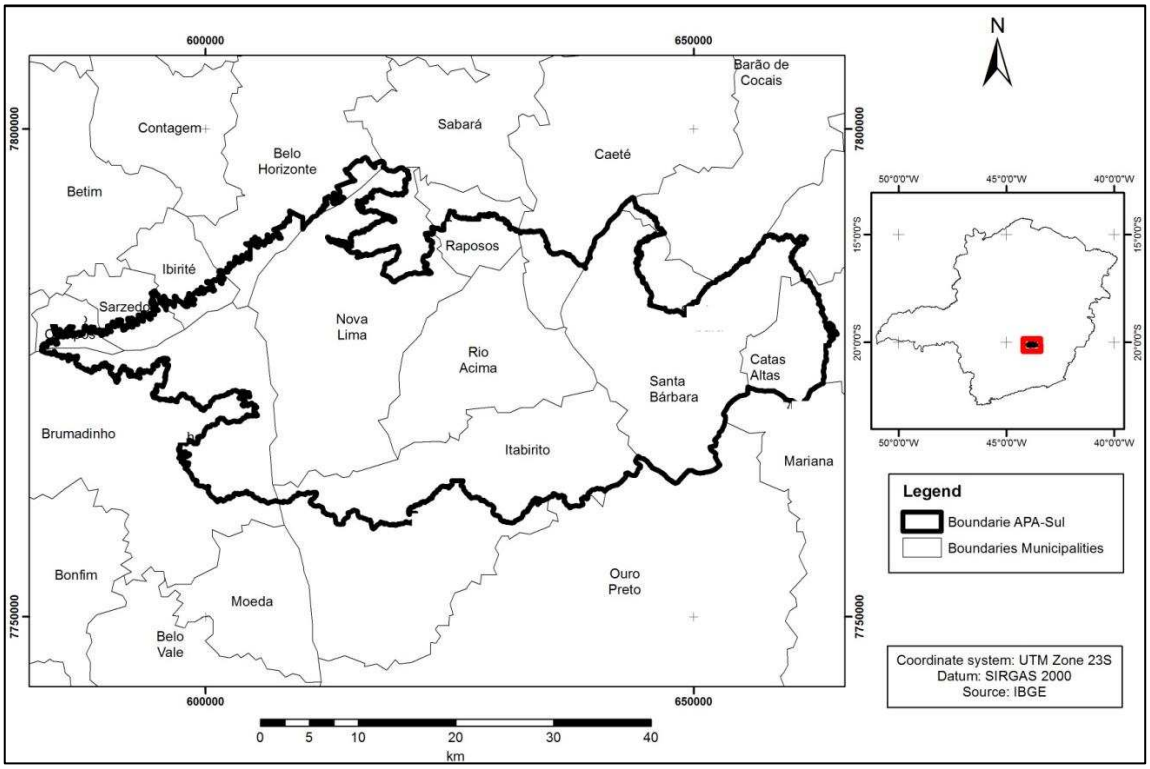


Figura 1. Localização da área de estudo.

A APA-Sul compreende uma área total de 1625,32 km^2, o clima da região é do tipo Cwa classificação de Köppen, temperado quente, com estação seca de abril a setembro e chuvosa de outubro a março. A região é caracterizada por um relevo estrutural emoldurado por um arcabouço geológico extremamente heterogêneo, constituído por duas principais unidades litoestratigráficas que são o Super Grupo Rio das Velhas e o Supergrupo Minas, ambas constituídas basicamente de rochas metamórficas do Pré-Cambriano

(CPRM, 2005). Em virtude desta diversidade geológica resulta uma cobertura pedológica também diversificada onde predominam solos poucos desenvolvidos como Cambissolos Háplicos e Neossolos Litólicos nas porções mais elevadas do relevo. Nas partes mais baixas da paisagem, encaixadas entre as cristas escarpadas, encontram-se solos um pouco mais desenvolvidos e profundos como Latossolos e Argissolos, desenvolvidos sobre rochas máficas intrusivas, ou sobre sedimentos coluvionares depositados nos sopés de algumas encostas. Próximo às regiões de drenagem encontram-se Neossolos Flúvicos e Gleissolos Háplicos. A cobertura vegetal da região apresenta fitofisionomias variando entre florestas estacionais semi decíduas, cerrados stricto sensu, campo cerrado e campos rupestres.

2.2. Solos: Mapas de referências e amostragem

Utilizou-se como referência um mapa de solos na escala 1:50.000 (CPRM, 2005) constituído por 75 unidades de mapeamento (UMs). Na área predominam solos pouco desenvolvidos como Neossolos Litólicos e Cambissolos Háplicos, ocorrendo também Latossolos, Argissolos, e com menor expressividade Gleissolos e Neossolos Flúvicos conforme descrito na Tabela 1.

Com vistas a avaliar a classificação em diferentes níveis de detalhamento, optou-se por simplificar as UMs desse mapa, gerando outros três mapas de solos compostos por 9, 24 e 34 UMs de solos, como descrito a seguir. O menor agrupamento ficou resumido ao segundo nível categórico, do primeiro componente das unidades de mapeamento, gerando assim um mapa com apenas 9 UMs. Os agrupamentos intermediários que geraram mapas com 24 e 34 UMs, foram elaborados a partir de uma simplificação que levou em consideração as similaridades entre fases de relevo e componentes de associação das UMs, descritos respectivamente nas Tabelas 2 e 3. O quarto mapa foi o conjunto completo com as 75 unidades de mapeamento sem nenhuma simplificação. Portanto a partir desta etapa passou-se a trabalhar com quatro mapas de solo, um para cada um dos respectivos agrupamentos de UMs (9, 24,34 e 75).

Tabela 1. Unidades de Mapeamento de solos do mapa utilizado como referência

Siglas das UMs	Descrição das UMs
AR1	Afloramento de Rocha, forte ondulado
AR2	Afloramento de Rocha + Neossolo Litólico Distrófico típico, média cascalhenta, ambos montanhoso
AR3	Afloramento de Rocha + Cambissolo Háplico Tb distrófico léptico ou lítico, média muito cascalhenta, ondulado
AR4	Afloramento de Rocha, montanhoso + Cambissolo Háplico Tb distrófico léptico ou lítico, média muito cascalhenta, ondulado
AR5	Afloramento de Rocha + Neossolo Litólico Distrófico típico, média cascalhenta, ambos montanhoso
AR6	Afloramento de Rocha + Neossolo Litólico Psamítico típico, arenosa cascalhenta, ambos forte ondulado
AR7	Afloramento de Rocha + Neossolo Litólico Psamítico típico, arenosa cascalhenta, ambos suave ondulado
AR8	Afloramento de Rocha + Neossolo Litólico Psamítico típico, arenosa cascalhenta, ambos forte ondulado
AR9	Afloramento de Rocha + Neossolo Litólico Psamítico típico, arenosa cascalhenta, ambos montanhoso
AR10	Afloramento de Rocha, escarpado + Neossolo Litólico Distrófico típico, média cascalhenta, montanhoso
EC1	Exposição de Canga, forte ondulado
EC2	Exposição de Canga + Plintossolo Pétrico Litoplíntico perférrico, média muito cascalhenta, ambos montanhoso
CXbd1	Cambissolo Háplico Tb distrófico típico, média + Cambissolo Háplico Tb distrófico típico, média ou média cascalhenta, ondulado
CXbd2	Cambissolo Háplico Tb distrófico típico, média ou média cascalhenta + Neossolo Litólito Distrófico típico, muito cascalhenta, ondulado
CXbd3	Cambissolo Háplico Tb distrófico típico, média ou média pouco cascalhenta, ondulado
CXbd4	Cambissolo Háplico Tb distrófico típico, média ou argilosa endopedregosa, ondulado
CXbd5	Cambissolo Háplico Tb distrófico típico, média cascalhenta, montanhoso
CXbd6	Cambissolo Háplico Tb distrófico típico, média ou média cascalhenta, montanhoso

Continua…

Tabela 1. Continuação…

Siglas das UMs	Descrição das UMs
CXbd7	Cambissolo Háplico Tb distrófico típico, média ou média cascalhenta + Cambissolo Háplico Tb Distrófico latossólico, argilosa, ondulado
CXbd8	Cambissolo Háplico Tb distrófico típico, média muito cascalhenta + Cambissolo Háplico Tb distrófico latossólico argilosa, montanhoso
CXbd9	Cambissolo Háplico Tb distrófico típico, média cascalhenta + Latossolo Vermelho Distrófico típio ou câmbico argilosa, forte ondulado
CXbd10	Cambissolo Háplico Tb distrófico típico, média cascalhenta + Latossolo Vermelho Distrófico câmbico argilosa, forte ondulado
CXbd11	Cambissolo Háplico Tb distrófico típico, argilosa + Latossolo Vermelho Distrófico típico ou câmbico argilosa, forte ondulado
CXbd12	Cambissolo Háplico Tb distrófico argissólico, argilosa, forte ondulado
CXbd13	Cambissolo Háplico Tb distrófico léptico ou típico, média cascalhenta, montanhoso
CXbd14	Cambissolo Háplico Tb distrófico flúvico, argilosa, plano + Argissolo Vermelho Distrófico câmbico, argilosa, suave ondulado
CXbd15	Cambissolo Háplico Tb distrófico típico, média ou média argilosa, ondulado
CXbd16	Cambissolo Háplico Tb distrófico típico + Cambissolo Háplico Tb Distrófico léptico, média muito cascalhenta, forte ondulado
CXbd17	Cambissolo Háplico Tb distrófico léptico ou lítico, média cascalhenta + Neossolo Litólito Distrófico típico, média cascalhenta, ondulado
CXbd18	Cambissolo Háplico Tb distrófico típico, média ou média cascalhenta, ondulado
CXbd19	Cambissolo Háplico Tb distrófico léptico ou lítico+ Neossolo Litólito Distrófico típico, ambosmédia muito cascalhenta, montanhoso
CXbd20	Cambissolo Háplico Tb distrófico típico, média, ondulado
CXbd21	Cambissolo Háplico Tb distrófico típico, média ou média cascalhenta , ondulado
CXbd22	Cambissolo Háplico Tb distrófico típico + Cambissolo Háplico Tb Distrófico lítico,ambos média cascalhenta, forte ondulado
CXbd23	Cambissolo Háplico Tb distrófico típico + Cambissolo Háplico Tb Distroférrico léptico,ambos argilosa, forte ondulado
CXbd24	Cambissolo Háplico Tb distrófico léptico ou lítico + Neossolo Litólico Distrófico típico, ambos média muito cascalhenta, ondulado
CXbd25	Cambissolo Háplico Tb distrófico léptico ou lítico + Neossolo Litólico Distrófico típico, ambos média muito cascalhenta, forte ondulado

Continua…

Tabela 1. Continuação…

Siglas das UMs	Descrição das UMs
CXj1	Cambissolo Háplico Perférrico típico, média cascalhenta, montanhoso
CXj2	Cambissolo Háplico Perférrico típico, média pouco cascalhenta + Latossolo Vermelho Perférrico típico argilosa, ondulado
CXj3	Cambissolo Háplico Perférrico típico, média pouco cascalhenta + Latossolo Vermelho Perférrico câmbico argilosa, ondulado
CXj4	Cambissolo Háplico Perférrico típico, média pouco cascalhenta, forte ondulado + Afloramento de Rocha, montanhoso
CXj5	Cambissolo Háplico Perférrico petroplíntico, média muito cascalhenta + Latossolo Vermelho Perférrico petroplíntico, arg, ondulado
CXj6	Cambissolo Háplico Perférrico petroplíntico + Neossolo Litólico Distrófico típico,ambos média cascalhenta, suave ondulado
GXbd	Gleissolo Háplico Tb Distrófico + Gleissolo Melânico Tb Distrófico, ambos média ou arenosa, plano
RYbd	Neossolo Flúvico Distrófico típico, média ou média argilosa, plano
LVAw	Latossolo Vermelho Amarelo Ácrico ou típico + Cambissolo Háplico Distrófico típico, ambos argilosa, forte ondulado
LVAj	Latossolo Vermelho Amarelo Perférrico câmbico + Cambissolo Háplico Distroférrico típico, ambos argilosa, ondulado
LVd1	LATOSSOLO VERMELHO Distrófico típico, textura argilosa, A moderado, fase floresta tropical subperenifólia, relevo suave ondulado
LVd2	LATOSSOLO VERMELHO Distrófico típico, textura argilosa ou muito argilosa, A moderado, fase floresta tropical subperenifólia, relevo suave ondulado e ondulado
LVd3	LATOSSOLO VERMELHO Distrófico típico, textura argilosa ou muito argilosa, fase relevo forte ondulado e montanhoso + ARGISSOLO VERMELHO-AMARELO Distrófico típico, textura argilosa/muito argilosa, fase relevo forte ondulado, ambos com A moderado, fase floresta tropical subperenifólia
LVd4	LATOSSOLO VERMELHO Distrófico típico + CAMBISSOLO HÁPLICO Tb Distrófico típico, ambos com textura argilosa, A moderado, fase floresta tropical subperenifólia, relevo forte ondulado e ondulado
LVd5	LATOSSOLO VERMELHO Distrófico típico + LATOSSOLO VERMELHOAMARELO Distrófico típico, ambos de textura argilosa ou muito argilosa, A moderado, fase floresta tropical subperenifólia, relevo suave ondulado e ondulado
LVj1	Latossolo Vermelho Perférrico típico + Cambissolo Háplico Distroférrico típico, ambos argilosa, forte ondulado
LVj2	Latossolo Vermelho Perférrico típico, argilosa ou muito argilosa, forte ondulado

Continua…

Tabela 1. Continuação…

Siglas das UMs	Descrição das UMs
LVj3	Latossolo Vermelho Perférrico típico, argilosa ou muito argilosa + Cambissolo Háplico Distroférrico petroplíntico, média muito cascalhenta, suave ondulado
LVwf	Latossolo Vermelho Acriférrico argissólico, argilosa, forte ondulado
LVw	Latossolo Vermelho Acrico típico, argilosa, suave ondulado
PVAd1	Argissolo Vermelho Amarelo Distófico típico ou câmbico + Cambissolo Háplico Tb Distrófico, ambos argilosa, fort. Ond.
PVAd2	Argissolo Vermelho Amarelo Distófico típico ou câmbico + Cambissolo Háplico Tb Distrófico, ambos argilosa, ondulado
PVd2	Argissolo Vermelho Distófico típico ou câmbico + Cambissolo Háplico Tb Distrófico, ambos argilosa, ondulado
RLd1	Neossolo Litólico Distrófico típico, média cascalhenta, montanhoso
RLd2	Neossolo Litólico Distrófico típico + Cambissolo Háplico Tb Distrófico, ambos média cascalhenta, escarpado
RLd3	Neossolo Litólico Distrófico típico, média cascalhenta + Afloramento de rocha, ambos escarpado
RLd4	Neossolo Litólico Distrófico típico + Cambissolo Háplico Tb Distrófico, léptico ou lítico ambos média muito cascalhenta, montanhoso
RLd5	Neossolo Litólico Distrófico típico, média cascalhenta, ondulado
RLd6	Neossolo Litólico Distrófico típico, média cascalhenta, montanhoso
RLd7	Neossolo Litólico Distrófico típico, média muito cascalhenta + Neossolo Regolítico Distrófico típico, argilosa, montanhoso
RLd8	Neossolo Litólico Distrófico típico, média cascalhenta, montanhoso + Afloramento de rocha, ambos escarpado
RLd9	Neossolo Litólico Distrófico típico, média cascalhenta + Afloramento de rocha, ambos escarpado
RLd10	Neossolo Litólico Distrófico típico + Cambissolo Háplico Perférrico típico, ambos média muito cascalhenta, montanhoso
RLd11	Neossolo Litólico Distrófico típico, média cascalhenta, escarpado
RLd12	Neossolo Litólico Distrófico típico + Cambissolo Háplico Perférrico léptico ou típico, ambos média cascalhenta, montanhoso
RLd13	Neossolo Litólico Distrófico típico + Cambissolo Háplico Perférrico léptico ou lítico, ambos média cascalhenta, montanhoso
RLd14	Neossolo Litólico Distrófico típico, média cascalhenta + Afloramento de rocha, ambos escarpado

Fonte: CPRM, 2005.

Tabela 2. Agrupamento das Unidades de Mapeamento para 34 Ums

Número	Sigla das UMs	Critérios de agrupamento
1	PVd 1 + PVd 2	Associação com PVA
2	PVAd 1 +PVAd 2	Associação com CXbd
3	CXj 1	
4	CXj 2+CXj 3	Associação com Lvj
5	CXj 4+6	Ambos petropl¡nticos associados com RLd e AR
6	CXj5	
7	CXbd 2, 3, 4, 5, 6, 12, 13	Sem associação
8	CXbd 1, 15, 18, 20, 21	Sem associação e com relevo ondulado
9	CXbd 7,8 , 16, 22, 23	Associação com CXbd
10	CXbd 9, 10, 11	Associação com LVAd
11	CXbd 17,19,24,25	Associação com RLd
12	CXbd 14	
13	GXbd	
14	LVj 1+LVj 3	Associação com CXj
15	LVj2	
16	LVAj	
17	LVd 1, 2, 5	Associação com LVA, relevo ondulado e suave ondulado
18	LVd 3	
19	LVd 4	
20	LVw	
21	LVwf	
22	LVAw	
23	RLd 1,5, 6, 7,	Sem associação
24	RLd 2, 4	Associação com CXbd
25	RLd 3, 8, 9	Associação com AR
26	RLd 10, 12, 13	Associação com Cxj
27	RLd 14	
28	RLd 11	
29	RYbd	
30	AR1	
31	AR 3 e 4	Associação com CXbd e RLd
32	AR 2, 5, 6, 7, 7, 9, 10	Associação só com RLd
33	EC1 e EC2	
34	não solo (mineração e área degradada)	

Tabela 3. Agrupamento das Unidades de Mapeamento para 24 Ums

Número	Sigla das UMs	Critérios de agrupamento
1	PVd 1+ PVd 2	Associação com PVA
2	PVAd 1 +PVAd 2	Associação com CXbd
3	CXj2 + CXj3 + CXj1	Associação com Lvj
4	CXj 4+ 5 +6	Ambos petroplínticos associados com RLd e AR
5	CXbd 1, 2, 3, 4, 5, 6, 7,8 ,12 ,13, 16, 22, 23	Sem associação ou associados com CXbd
6	CXbd 17,19,24, 25	Associação com Rld
7	CXbd 15, 18, 20,21	Sem associação com relevo ondulado e suave ondulado
8	CXbd 9, 10, 11	Associação com LVAd
9	GXbd	
10	LVj1 + LVj3	Associação com Cxj
11	LVj2 + LVAj	Sem associação
12	LVd1, 2, 5	Sem associação ou associado com LVA relevo ondulado e suave ondulado
13	LVd3 e 4	Ambos relevo forte ondulado porém com associações diferentes
14	LVw + LVAw + Lvwf	Ambos ácricos
15	RLd 1,5, 6, 7,	Sem associação
16	RLd 2, 4	Associação com CXbd
17	RLd 3, 8, 9	Associação com AR
18	RLd 10, 12, 13	Associação com CXj
19	RLd 14	Associação com AR
20	RLd 11	Sem associação
21	RYbd + CXbd 14	relevo plano/ondulado associado com PVA
22	AR 3 e 4	Associação com CXBbd e RLd
23	AR1, 2, 5, 6, 7, 8, 9, 10 + EC1 e EC2	Associação só com RLd e AR 1 que não tem associação
24	não solo (mineração e área degradada)	

As amostras utilizadas para treinamento do classificador Random Forest, e de validação a partir da validação cruzada, foram obtidas a partir do mapa de referência (CPRM, 2005), por amostragem aleatória. A densidade de pontos amostrais foi definida a partir do critério indicativo do número de observações para mapeamentos em escala semidetalhada presente no manual de procedimentos normativos de levantamentos pedológicos (EMBRAPA, 1995) e no manual técnico de pedologia (IBGE, 2015). Adotou-se uma densidade de amostragem de uma observação a cada 5 ha, que aplicada na área de 162.503 ha resultou em 32.500 pontos. Apesar de este valor ser elevado e possivelmente de caráter impraticável em campo, foi adotado porque o objetivo era avaliar o desempenho do conjunto de variáveis em condições ótimas.

2.3. Covariáveis preditivas

As covariáveis utilizadas para a predição, agrupadas de acordo com os fatores de formação de solos associados estão apresentadas no Tabela 4. As covariáveis morfométricas foram obtidos do MDE gerado com imagens SRTM (Shuttle Radar Topography Mission) com 90 metros de resolução (CGIAR, 2014). Além disso, utilizou-se dois mapas temáticos categóricos na escala de 1:50.000, com informações de relevo, sendo um de unidades de relevo e o outro de geomorfologia (CPRM, 2005). As variáveis morfométicas, derivadas do MDE, foram geradas através do software R com o pacote R-Saga. As variáveis espectrais que, relacionam-se ao fator organismos, foram obtidas de imagem do satélite Landsat 8. Referente ao fator material de origem, utilizou-se os mapas temáticos de geologia 1:25.000 (LOBATO et al., 2005), e os dados aerogeofísicos de gamaespectometria, na forma de três canais (Th,U, K) e as respectivas razões (Th/K, U/Th, U/K), com resolução original de 125 m. Os dados geofísicos provêm de técnicas indiretas (Sensoriamento Remoto) de investigação das estruturas em subsuperfície. Estes dados permitem estimar as condições geológicas através do contraste das propriedades físicas dos materiais de subsuperfície. O levantamento gamaespectrométrico reflete variações geoquímicas do K, U e Th, que são os únicos elementos de ocorrência natural com radioisótopos capazes de produzir raios gama com suficiente energia para serem dectadas por receptores. A média crustal destes

elementos é de 2,31% para o K; 2,7 ppm para U e 10,5 ppm para o Th (Minty, 1997).

Tabela 4. Covariáveis utilizadas, subdivididas pelos fatores de formação de solos

Relevo		**Clima**	
Aspect	Slope	BIO1=Annual Mean Temperature	Maximum, mean and minimum monthly temperatures
Aspect reclassified	MRRTF (Multi-resolution ridge top flatness)	BIO2=Mean Diurnal Range (Mean of monthly (max temp-min temp))	
Convergence index	MRVBF (Multi-resolution valley bottom flatness)	BIO3=Isothermality (BIO2/BIO7)(*100)	Monthly precipitations
Cross secational curvature	Slope Height	BIO4=Temperature Seasonality (standarddeviation*100)	
Flow line curvature	Mass balance index	BIO4=Temperature Seasonality (standarddeviation*100)	
General curvature	Índice de vale	BIO5=Max Temperature of WarmestMonth	
Longitudinal curvature	Solar radiation diffuse 1	BIO6=MinTemperature of ColdestMonth	
Maximal curvature	Solar radiation diffuse 2	BIO7=Temperature Annual Range (BIO5-BIO6)	
Minimal curvature	Solar radiation direct 1	BIO8=Mean Temperature of Wettest Quarter	
Plan curvature	Solar radiation direct 2	BIO9=Mean Temperature of Driest Quarter	
Profile curvature	Solar radiation duration 1	BIO10=Mean Temperature of Warmest Quarter	
Tangencial curvature	Solar radiation duration 2	BIO11=Mean Temperature of Coldest Quarter	

Continua...

Tabela 4. Continuação...

Relevo		Clima	
Total curvature	Solar radiation total 1	BIO12=Annual Precipitation	
Diferençe	Solar radiation total 2	BIO13=Precipitation of Wettest Month	
Euclidean distance drainage	Surface specific points	BIO14=Precipitation of Driest Month	
Diurnal anysotropic heating	Terrain Ruggednes Index	BIO15=Precipitation Seasonality (Coefficient of Variation)	
Gradient	Terrain Surface convexity	BIO16=Precipitation of Wettest Quarter	
Hill index	Terrain Surface Texture	BIO17=Precipitation of Driest Quarter	
Landforms	Topographic Position Index	BIO18=Precipitation of Warmest Quarter	
Standardized height	Topographic Wetness Index (TWI)	BIO19 = Precipitation of Coldest Quarter	
MDE	Valley Index	Maximum, mean and minimum temperature in month 1	
Mid slope position	Valley	Maximum,mean and minimum temperature in month 2	
Morphometric protection index	Valley Depth	Maximum,mean and minimum temperature in month 3	
Normalized height	Vector Ruggedness Measure		
Real surface area	Relief Units Map		
Slope reclassify	Geomorphologic Units Map		
WTI (1)			

Continua...

Tabela 4. Continuação...

Organismos	Material de Origem
Band 1 of Landsat 8	Geology map (1:25.000)
Band 2 of Landsat 8	Clay minerals index
Band 3 of Landsat 8	Iron oxides index
Band 4 of Landsat 8	Iron index
Band 5 of Landsat 8	Concentration of the K element (gammaspectrometry)
Band 6 of Landsat 8	Concentration of the Th element (gammaspectrometry)
Band 7 of Landsat 8	Concentration of the U element (gammaspectrometry)
NDVI	Ratio between the concentrations of Th / K (Gammaspectrometry)
	Ratio between the concentrations of Th / U (Gammaspectrometry)
	Ratio between the concentrations of U / K (Gammaspectrometry)Band 2 magnetometry
	Band 2 magnetometry

A detecção dos raios gama só ocorreu após a década de 1950, visto que até então não existiam detectores capazes de distinguir os raios gama, gerando somente uma contagem total da energia radioativa. Após o surgimento do detector feito a partir de iodeto de sódio (NaI) foi possível individualizar o registro dos pulsos referentes à energia da radiação gama (MINTY, 1997). Portanto a detecção baseia-se na interação das partículas com o material sensível do detector, sendo os pulsos gerados amplificados e discriminados no analisador multicanal, que os armazena em canais distintos determinados pelas diferenças no comprimento de onda dos elementos. A quantificação dos teores dos elementos é mais ou menos exata, a depender do tipo de equipamento (detector) e da altura do vôo do levantamento. Adota-se como padrão a altura de vôo de 100 m.

As partículas gama constituem a parcela mais energética das substâncias radioativas, podendo atravessar até 30 cm superficiais da crosta. Segundo Wilford et al. (1997), durante o intemperismo os radioelementos são liberados da rocha e redistribuídos no saprolito e ou solo, portanto essa camada pode apresentar perda de K e em muitos casos acúmulo relativo de Th e U. Ainda conforme estes autores as rochas intermediárias e básicas mostram pouca mudança dos radioelementos durante o intemperismo inicial. Porém com a pedogênese podem ser gerados solos com duas ou três vezes mais U e Th que a rocha de origem. Desta forma os solos apresentam radioatividade diretamente relacionada ao material de origem, bem como aos processos pedogenéticos vigentes.

Por fim as variáveis associadas ao clima foram obtidas da base de dados WorldClim – Global Climate Data (WORLDCLIM, 2015) com resolução original de 1000 m. Foram acrescidas ao conjunto total de variáveis preditivas, as coordenadas geográficas geradas em sistema de coordenadas Universal Transversa de Mercator (UTM).

2.4. Predição dos mapas

O Random Forest (RF) foi utilizado por ser um classificador que apresenta bom desempenho em predição de classes, e pelo fato possuir um ajuste de modelo relativamente simples, sendo neste estudo ajustado os parâmetros mtry = 4 e ntree = 1000. O RF foi desenvolvido como uma extensão dos modelos de Árvores de classificação. Trata-se de uma combinação de árvores preditoras que dependem dos valores de um vetor aleatório amostrado independentemente mas com a mesma distribuição para todas as árvores (BREIMAN, 2001). Dentre as vantagens deste método destaca-se que ele trabalha conjuntamente com variáveis categóricas e numéricas, apresenta grande eficiência e rapidez no processo de treinamento e possui três parâmetros a serem ajustados, sendo estes o número de variáveis por nó (mtry), o número de árvores do modelo (ntree) e o número de nós na ramificação final das árvores. Mas a vantagem essencial do RF foi o ranqueamento das variáveis por ordem de importância. Issa ordem é gerada a partir do sistema de estimativa de erro interno do RF chamado Out-Of-Bag. O RF altera o arranjo das variáveis no conjunto Out-Of-Bag, e estima a

importância da variável através da variação causada no erro (TAGHIZADEH-MERJARDHI, 2016).

2.5. Seleção de covariáveis

Neste estudo foram avaliados quatro procedimentos de seleção de covariáveis, sendo dois elaborados e propostos neste estudo (Percentil e Correlação) e outros dois (Índice Ginni e Mean Decrease Accuracy MDA) baseados essencialmente no ranqueamento de covariáveis gerado pelo comando Importance do algorítimo Random Forest. Portanto, buscou-se comparar o comportamento dos métodos propostos frente a métodos já comumemente utilizados.

Os valores de índice kappa obtidos para cada um dos métodos foram comparados pelo Teste Z a 5% de probabilidade, permitindo assim estabelecer os pontos de corte, onde existia diferença estatística dos maiores para os menores valores deste índice. Foram gerados também os valores do índice kappa condicional, permitindo assim avaliar o desemepenho do treinamento para cada classe existente nos quatro agrupamentos trabalhados (9, 24, 34 e 75 UMs).

2.5.1. Seleção baseada no ranqueamento do Random Forest

Os métodos doÍndice Ginni e Mean decrease Accuracy (MDA), utilizam a lista de covariáveis apresentada pelo ranqueamento do RandomForest, selecionadas pelo Índice Ginni e pelo MDA respectivamente. Em cada uma das duas listas ordenadas, aplicou-se a remoção de variáveis, uma a uma, da menor para a maior importância em cada ranqueamento, obtendo-se também o índice kappa para cada um dos conjuntos de variáveis compostos por no maximo 141 (conjunto total) e mínimo 2 covariáveis.

2.5.2. Proposta de seleção de covariáveis

Os dois procedimentos aqui propostos são bem distintos metodológicamente. O baseado na variação do Percentil trata-se de um

procedimento híbrido que utiliza o Percentil juntamente com o ranqueamento de covariáveis do Ramdon Forest. O percentil trata-se de uma subdivisão da amostra ordenada por ordem crescente, em 100 partes com porcentagens aproximadamente iguais.

A aplicação do procedimento foi feita em cima de uma média obtida a partir de 10 repetições, que consistiram de 10 calssificações feitas com o Ramdon Forest, onde se utilizou o comando Importance para selecionar as covariáveis mais importantes para cada classe a ser predita. A partir destas covariáveis pré selecionadas o Percentil foi aplicado como o direcionador da remoção de covariáveis menos importantes. Para tanto foi feito uma classificação para cada valor de Percentil, variando de 1 em 1, no intervalo de 1 a 99. Por fim, obteve-se um valor de índice kappa para cada percentil, permitindo assim avaliar o comportamento do índice kappa a medida que uma certa porcentagem de covariáveis vai sendo removida.

O outro procedimento consiste na remoção de covariáveis com alta correlação não linear. Iniciou-se com a remoção das covariáveis mais correlacionadas, em um intervalo de 10% de correlação. Posteriormente, foi realizado um detalhamento entre 30% a 60% de correlação onde o intervalo aplicado foi de 1%, visando o melhor detalhamento do ponto de inflexão da curva. Este procedimento baseia-se na remoção da sobreposição de informação, visto que variáveis altamente correlacionadas carregam informações semelhantes.

2.6. Validação (compação pixel a pixel entre os mapas preditos e os mapas de referências)

Com o intuito de validar definitivamente os conjuntos de variáveis elencados no processo de treinamento do classificador RF, realizou-se a comparação entre cada um dos mapas preditos nos quatro agrupamentos de classes (9, 24, 34 e 75) e os respectivos mapas de referência. A exatidão dos mapas preditos foi avaliada através da matriz de erro, onde se tem nas colunas os dados de referência e nas linhas os dados classificados, sendo a diagonal principal a representação do nível de concordância entre ambos mapas (CONGALTON e GREEN, 1999). No entanto para compensar os acertos ao acaso que não são levados em consideração pela matriz de confusão optou-se

por utilizar o Índice Kappa que trata-se de uma estatística multivariada discreta utilizada para medir a concordância entre dados estimados e de referência mas que desconsidera os acertos ocorridos ao acaso. O Índice Kappa normalmente varia entre 0 e 1, sendo 0 ausência de concordância e 1 concordância total (CONGALTON e GREEN, 1999).

3. RESULTADOS

3.1. Comparação entre os métodos de seleção de variáveis

Os valores do índice Kappa e do Accuracy (exatidão glogal) obtidos para cada um dos métodos de seleção de variáveis estão apresentados nas Figuras 2, 3, 4 e 5, sendo cada uma associada a um dos quatro agrupamentos de classes (9, 24, 34, 75) respectivamente. Os quatro métodos possuem um patamar bem definido onde os valores dos índices de kappa não se distinguem estatisticamentepelo Teste Z a 5%. No entanto, a medida que as variáveis vão sendo removidas observa-se uma diferença expressiva na posição em que se encontra o ponto onde os valores começam a decrescer.

Através da análise gráfica das Figuras 2 a 5, é possível perceber que houve pequena diferença entre a seleção de variáveis feita através do ranqueamento fornecido pelo índice de Ginni e pelo MeanDecrease Accuracy. Ambos possuem um patamar bem estabilizado e uma queda brusca nos valores do índice Kappa próximo ao conjunto com seis variáveis. Na remoção de variáveis feita através do percentil, também observa-se um patamar bem estável até próximo do valor 0,97 de percentil, evidenciando que o classificador manteve o padrão de exatidão alto, utilizando menos de 3% das variáveis mais importantes. Já nos gráficos de remoção de variáveis por porcentagem da correlação, nota-se que o decréscimo mais expressivo nos valores de acurácia está próximo a 35% de correlação. Portanto o classificador apresentou perda de acurácia quando variáveis com menos de 35% de correlação começaram a ser removidas.

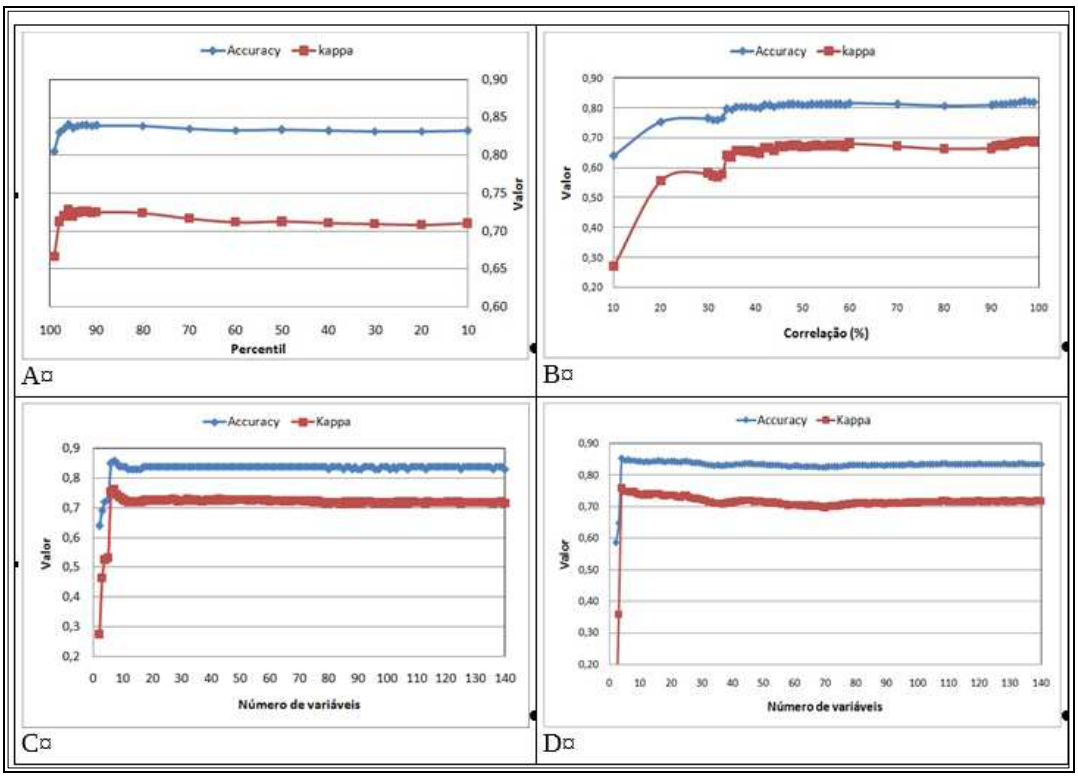


Figura 2. Comparação entre os resultados de Kappa e Meandecrease Accuracy para o mapa de 9 classes de solo, com os procedimentos de seleção de variáveis: A- Percentil; B- Correlação; C- Índice Ginni; D- Mean Decrease Accuracy.

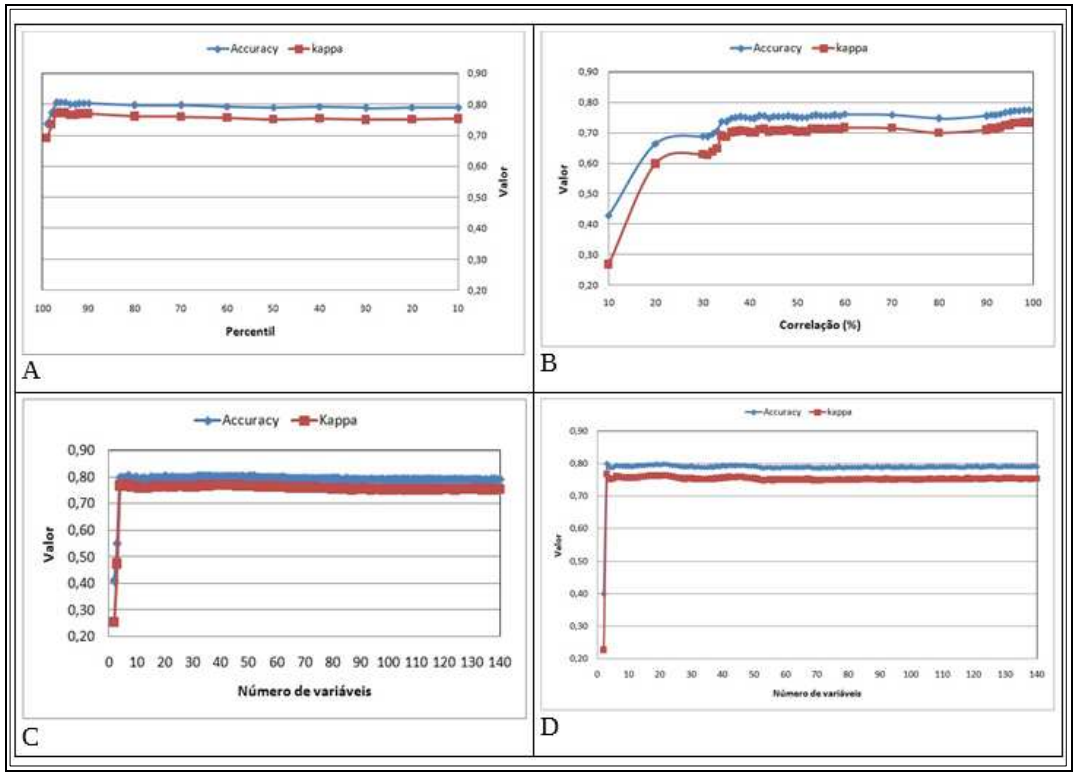


Figura 3. Comparação entre os resultados de Kappa e Meandecrease Accuracy para o mapa de 24 classes de solo, com os procedimentos de seleção de variáveis: A- Percentil; B- Correlação; C- Índice Ginni; D- Mean Decrease Accuracy.

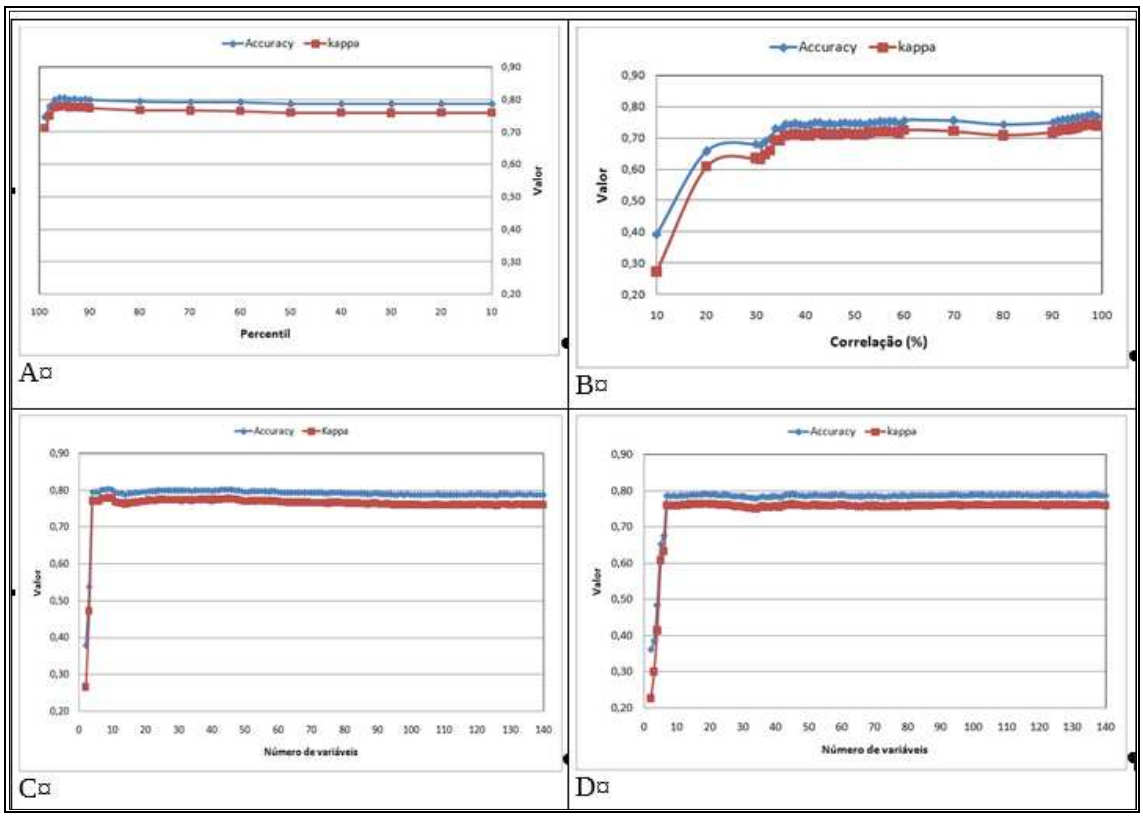


Figura 4. Comparação entre os resultados de Kappa e Meandecrease Accuracy para o mapa de 34 classes de solo, com os procedimentos de seleção de variáveis: A- Percentil; B- Correlação; C- Índice Ginni; D- Mean Decrease Accuracy.

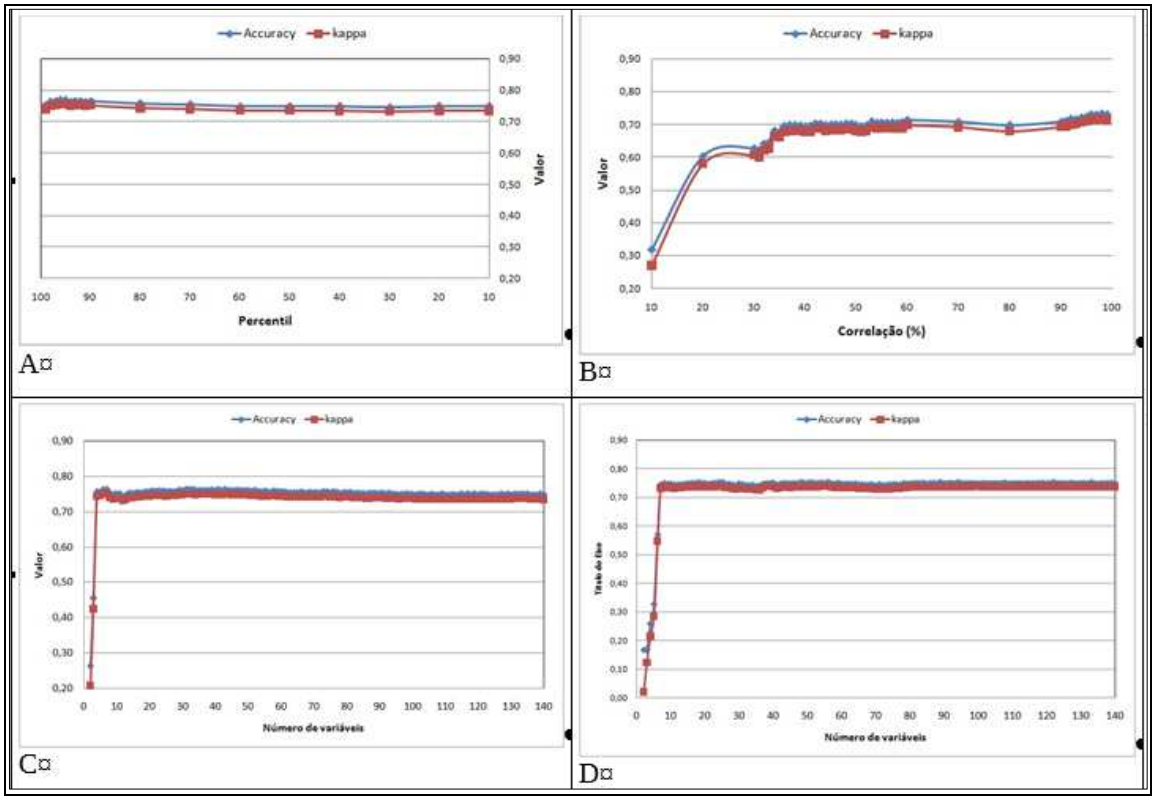


Figura 5. Comparação entre os resultados de Kappa e Meandecrease Accuracy para o mapa de 75 classes de solo, com os procedimentos de seleção de variáveis: A- Percentil; B- Correlação; C- Índice Ginni; D- Mean Decrease Accuracy.

Aplicando-se o teste Z a 5% foi possível estabelecer o ponto de corte, onde os valores de exatidão são menores e diferentes estatisticamente do patamar estabelecido por valores melhores. O ponto de corte representa o melhor valor de índice Kappa onde se tenha o menor número de variáveis e onde todas as classes são contempladas pelo classificador.

Na Tabela 5 estão apresentados os pontos de corte estabelecidos para cada um dos métodos avaliados com os respectivos números de variáveis selecionadas, valores do índice kappa e número de UMs de solos não contempladas na seleção.

A comparação entre os quatro métodos de seleção de variáveis, pelo índice Kappa, mostra desempenho semelhante entre os métodos de Percentil, índice Ginni e Meandecrease Accuracy. Já o método da correlação apresentou desempenho inferior com valores de índice Kappa menores do que todos os outros três métodos. No entanto, quando esta análise é feita com relação ao número de variáveis selecionadas observa-se uma similaridade somente entre o índice Ginni e Meandecrease Accuracy, que obtiveram o menor número de variáveis. O método do Percentil selecionou um número intermediário de variáveis, e o método da correlação um número bem maior do que os demais.

Somente nos mapas de 75 UMs não se conseguiu predizer todas as UMs, variando o número de UMs não contempladas entre 7 e 4.

As Tabelas 6 e 7 mostram as variáveis selecionadas pelos procedimentos do Percentil, Índice Ginni e Meandecrease Accuracy para cada um dos agrupamentos de classes. Nota-se que algumas variáveis foram selecionadas como prioritárias em praticamente todos os ranqueamentos. Dentre estas destacam-se as coordenadas geográficas, e os mapas de relevo e de geomorfologia, que apareceram sempre entre as sextas ou sétimas variáveis mais importantes.

Tabela 5. Pontos de corte, índice Kappa, número de variáveis selecionadas e número de UMs não contempladas em cada um dos procedimentos testados para os quatro agrupamentos de solos

Nº de unidades de solo	Ponto de corte	Índice Kappa	Nº de variáveis	Nº de UMs não contempladas
---------- Percentil ----------				
9	0,98	0,71	8	0
24	0,97	0,77	16	0
34	0,97	0,77	18	0
75	0,98	0,75	14	7
---------- Índice Ginni ----------				
9	6	0,75	6	0
24	4	0,76	4	0
34	4	0,77	4	0
75	4	0,74	4	4
---------- Meandecrease Accuracy ----------				
9	4	0,76	4	0
24	3	0,76	3	0
34	7	0,76	7	0
75	7	0,73	7	6
---------- Correlação (%)----------				
9	60	0,68	34	0
24	60	0,71	34	0
34	60	0,72	34	0
75	60	0,69	34	7

Tabela 6. Variáveis selecionadas pelos procedimentos: Percentil, Índice Ginni e Meandecrease Accuracy para 9 e 24 UMs de solo

Percentil	**Índice Ginni**	**Meand. Accuracy**
9 UMs		
Map of Relief units	Map of Relief units	Longitude (Y)
Latitude (X)	Latitude (X)	Map of Relief units
Longitude (Y)	Geology map (1:25.000)	Latitude (X)
Geology map (1:25.000)	Geomorphologycal Map	Ratio between the concentrations of Th / K (Gammaspectrometry)
Geomorphologycal Map	Longitude (Y)	-
MDE	MDE	-
MRVBF	-	-
Standardized height	-	-
24 UMs		
Map of Relief units	Map of Relief units	Latitude (X)
Latitude (X)	Latitude (X)	Longitude (Y)
Longitude (Y)	Longitude (Y)	Terrain surface texture
Geomorphologycal Map	Geomorphologycal Map	-
Geology map (1:25.000)	-	-
Concentration of the K element (gammaspectrometry)	-	-
Concentration of the Th element (gammaspectrometry)	-	-
Precipitation in month 1	-	-
MDE	-	-
Slope Height	-	-
Standardized height	-	-
Valley depth	-	-
Vector Ruggedness Measure	-	-
Clay minerals index	-	

Tabela 7. Variáveis selecionadas pelos procedimentos: Percentil, Índice Ginni e Meandecrease Accuracy para 34 e 75 classes

Percentil	Índice Ginni	Meand. Accuracy
34 UMs		
Map of Relief units	Map of Relief units	Latitude (X)
Latitude (X)	Latitude (X)	Terrain surface texture
Geomorphologycal Map	Longitude (Y)	Slope Height
Longitude (Y)	Geomorphologycal Map	Map of Relief units
Geology map (1:25.000)	-	Standardized height
Concentration of the K element (gammaspectrometry)	-	Longitude (Y)
BIO2=Mean Diurnal Range (Mean of monthly (max temp-min temp))	-	Vector Ruggedness Measure
MDE	-	-
Precipitation in month 11	-	-
Concentration of the Th element (gammaspectrometry)	-	-
Slope Height	-	-
Terrain surface texture	-	-
Clay minerals index	-	-
Precipitation in month 1	-	-
Precipitation in month 4	-	-
Standardized height	-	-
(WTI-1)	-	-
75 UMs		
Map of Relief units	Latitude (X)	Terrain surface texture
Latitude (X)	Map of Relief units	Longitude (Y)
Longitude (Y)	Longitude (Y)	Slope Height
Geomorphologycal Map	Geomorphologycal Map	Vector Ruggedness Measure
Geology map (1:25.000)	-	Map of Relief units
MDE	-	Latitude (X)
BIO2=Mean Diurnal Range (Mean of monthly (max temp-min temp))	-	Valley depth
Precipitação no mês 4	-	-
Precipitação no mês 11	-	-
Concentration of the K element (gammaspectrometry)	-	-
Concentration of the Th element (gammaspectrometry)	-	-
Slope Height	-	-
Standardized height	-	-
Terrain surface texture	-	-

A variável altitude (MDE) apareceu sempre vinculada a seleção feita pelo método do percentil, por outro lado só apareceu uma vez na seleção do índice Ginni para 9 classes. O mapa geológico é outra variável importante que apesar de ter sido predominantemente selecionada pelo método do Percentil (Tabelas 6 e 7), sempre que aparece está entre as cinco variáveis mais importantes.

As demais variáveis selecionadas tiveram representação dos quatro fatores de formação de solos (relevo, organismos, clima e material de origem) aqui contemplados. Dentre as relacionadas ao relevo, destacam-se Terrain Surface Texture, Vector Ruggedness Measure, Standardized height e Slope Height. As variáveis geofísicas, vinculadas ao material de origem, aparecem em todos os conjuntos selecionados pelo percentil, menos para 9 UMs. Porém a razão Th/K foi selecionada para o conjunto com nove classes pelo Mean decrease Accuracy. Além disso, estas variáveis também apareceram com frequência no conjunto selecionado pelo método da correlação (Tabela 8).

O Percentil e a Correlação foram os únicos que selecionaram variáveis climáticas, com destaque para precipitação nos meses, 1, 4 e 11 e a Temperatura média diurna mensal (Bio 2). O procedimento baseado na Correlação foi o único a selecionar variáveis relacionadas aos organismos, destacando-se o NDVI e as bandas 4 e 5 da Landsat 8.

3.2. Comparação pixel a pixel entre os mapas preditos e os mapas de referências

Na Figura 7 são apresentados os valores de exatidão obtidos a partir da comparação entre os mapas preditos e os mapas de referência. Avaliando-se os valores do índice Kappa para cada um dos métodos observa-se que para o conjunto com 9 UMs de solos, o melhor valor de Kappa foi obtido para o mapa gerado com as variáveis selecionadas pelo índice Ginni. Nos demais agrupamentos de UMs os melhores valores de Kappa estão associados aos mapas elaborados com o conjunto de variáveis selecionado pelo método do percentil.

O conjunto de variáveis selecionado pelo método de correlação forneceu em todos os agrupamentos de classes os piores resultados, evidenciando a menor eficiência deste método.

Tabela 8. Variáveis selecionadas pelo método de remoção por Correlação para os agrupamentos de UMs (9, 24, 34 e 75)

9UMs	24 UMs	34 UMs	75 UMs
BIO2=Mean Diurnal Range (Mean of monthly (max temp-min temp))	Band 2 magnetometry	Gradient	Valley Index
BIO2=Mean Diurnal Range (Mean of monthly (max temp-min temp))	Ratio between the concentrations of Th / K (Gammaspectrometry)	MDE	Valley
Precipitation in month 10	Ratio between the concentrations of U/Th (Gammaspectrometry)	Morphometric protection	Band 4 of Landsat 8
Minimum temperature in month 9	Ratio between the concentrations of U / K (Gammaspectrometry)	MRRTF (Multi-resolution ridge top flatness)	NDVI
Geology map (1:25.000)	Cross secational curvature	Solar radiation direct 1	Banda 5 da Landsat 8
Geomorphologycal Map	Minimal curvature	Solar radiation diffuse 2	Aspect reclassify
Map of Relief units	Tangencial curvature	Terrain Ruggednes Index	Standardized height
Concentration of the K element (gammaspectrometry)	Total curvature	Terrais Surface convexity	
Concentration of the Th element (gammaspectrometry)	Difference	Topographic Position Index	

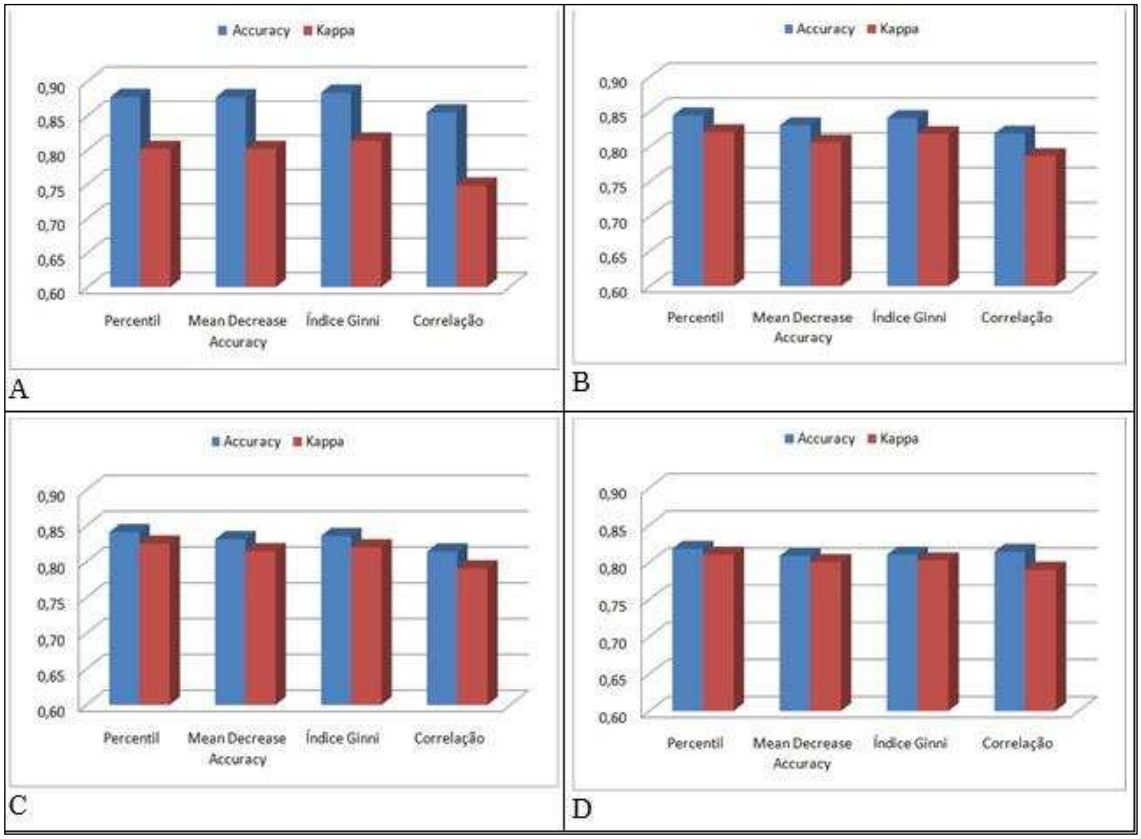


Figura 7. Valores de exatidão provenientes da comparação entre os mapas preditos e os mapas de referência: A- 9 UMs; B- 24 UMs; C- 34 UMs; D- 75 UMs.

Aplicou-se o Teste Z a 5% sobre os valores do índice kappa, e foi possível diagnosticar que entre os melhores resultados, o maior valor foi atribuído a predição do conjunto com 34 UMs, que apesar de muito próximo do valor obtido para 24 UMs, diferiu estatisticamente deste último. O menor valor entre os diferentes agrupamentos foi observado para 75 UMs. A Figura 8 evidencia graficamente as diferenças observadas. Desta forma o mapa de melhor exatidão, kappa = 0,82, foi obtido com as variáveis selecionadas pelo método do percentil para 34 UMs.

Nas Figuras 9, 10, 11 e 12 são apresentados os mapas de incerteza. Nestes mapas destaca-se somente as áreas onde houve discordância entre os mapas preditos, com o melhor conjunto de variáveis para os quatro agrupamentos de UMs de solos: 9, 24, 34 e 75 respectivamente, e os mapas de referência, gerados pelo agrupamento de UMs a partir do mapa convencional de solos. A análise destas figuras demonstra que os mapas produzidos apresentam uma distribuição aleatória dos erros ao longo de toda a área de estudo. No entanto, é possível perceber que as áreas de incerteza

estão intimamente associadas com as áreas limítrofes entre as UMs. Este fato é compreensível, visto que é no limite entre classes ou UMs que existem as maiores chances de erro. Portanto como a análise foi feita sobre um mapa discreto, onde as classes são individualizadas com limites abruptos, quando se faz uma análise mais quantitativa, onde a transição de uma classe para outra é gradativa torna-se mais evidente a chance de erros. Visualmente não é possível uma distinção criteriosa dos resultados obtidos, evidenciando um comportamento similar entre os mapas gerados, coincidente com as tênues diferenças entre os valores de índice kappa obtidos.

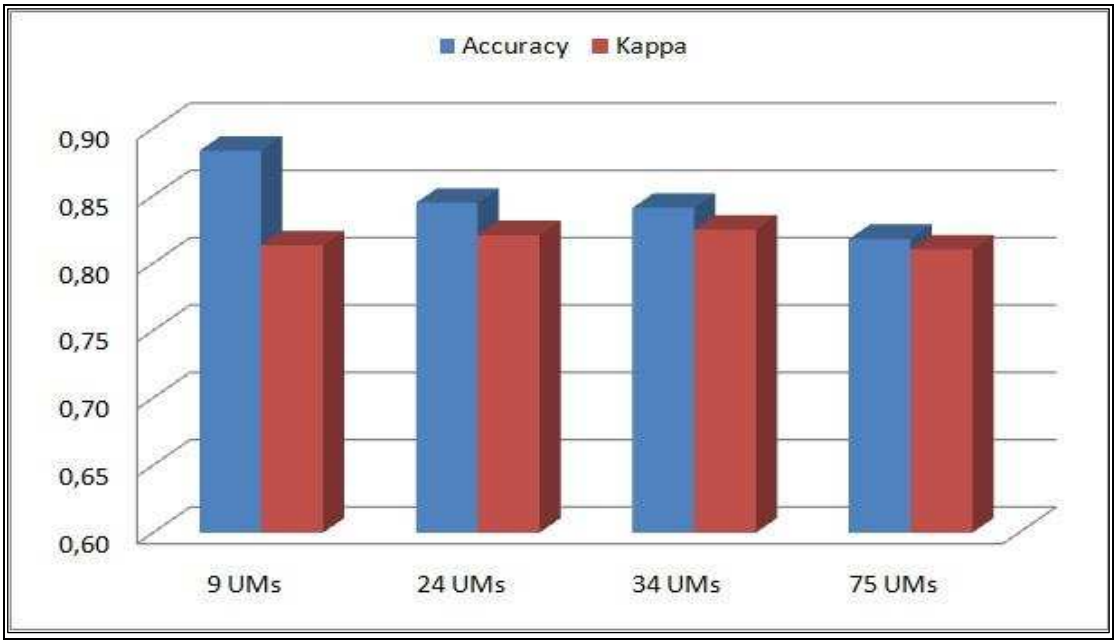


Figura 8. Comparação entre os melhores valores do índice Kappa obtidos para cada um dos agrupamentos de Unidades de Mapeamento (UMs).

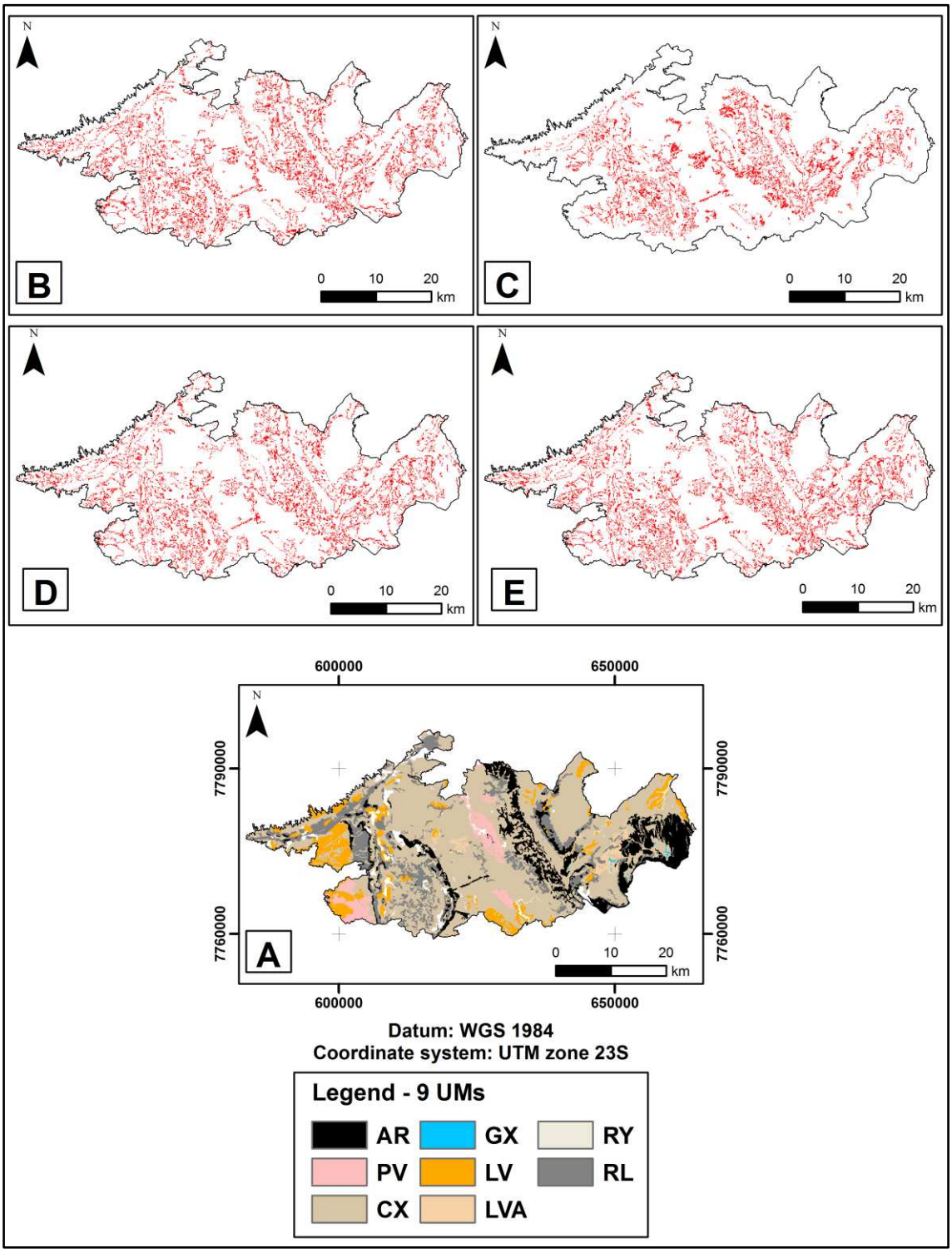


Figura 9. Mapas de incerteza para 9 UMs: discordância entre o mapa de referência e os mapas preditos pelos conjuntos de variáveis selecionados por cada um dos métodos avaliados: A- Mapa de solos de referência (CPRM, 2005); B- Método do Meandecrease Accuracy; C- Método da Correlação; D- Método do Índice Ginni; E- Método do Percentil.

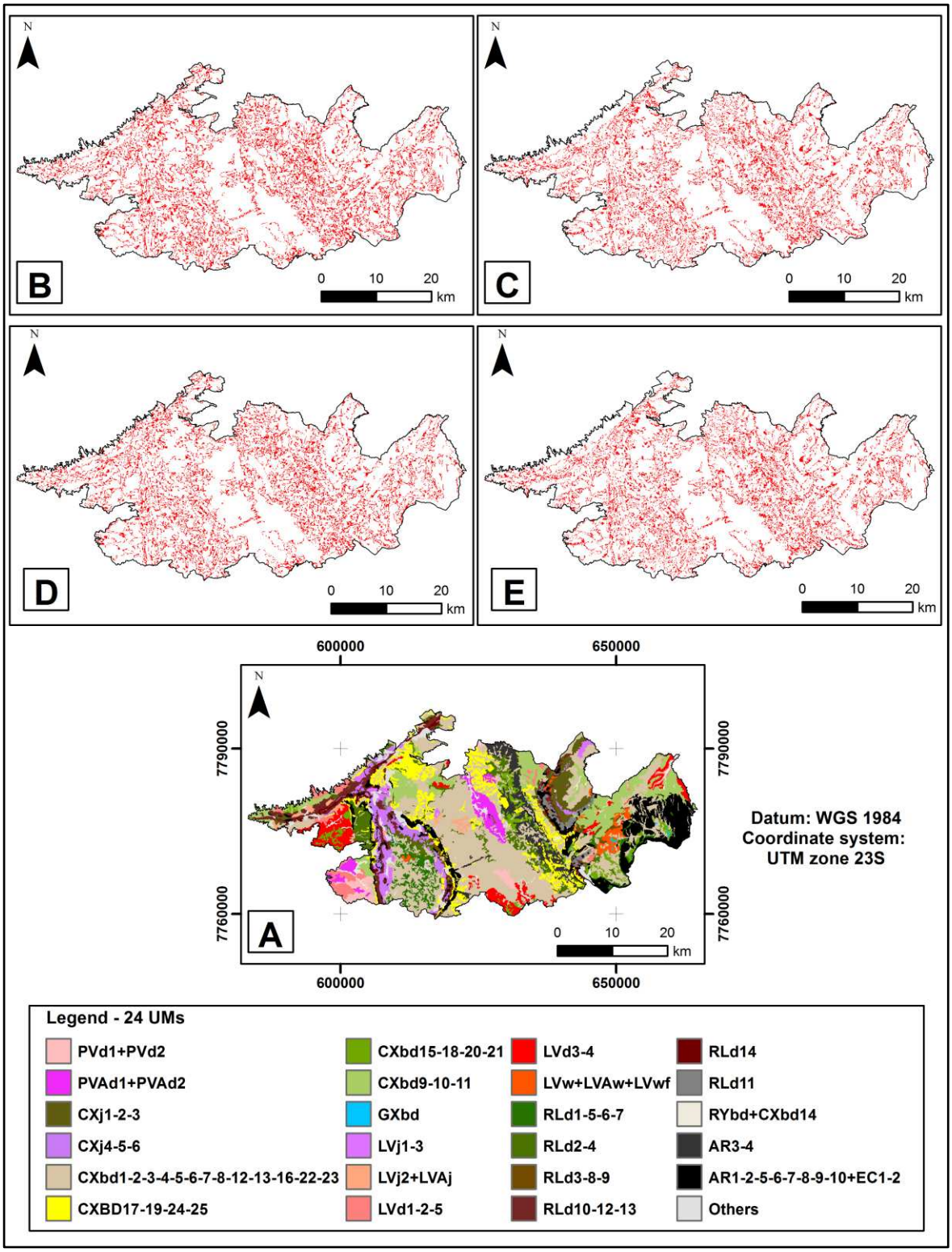


Figura 10. Mapas de incerteza para 24 UMs: discordância entre o mapa de referência e os mapas preditos pelos conjuntos de variáveis selecionados por cada um dos métodos avaliados: A- Mapa de solos de referência (CPRM, 2005); B- Método do Meandecrease Accuracy; C- Método da Correlação; D- Método do Índice Ginni; E- Método do Percentil.

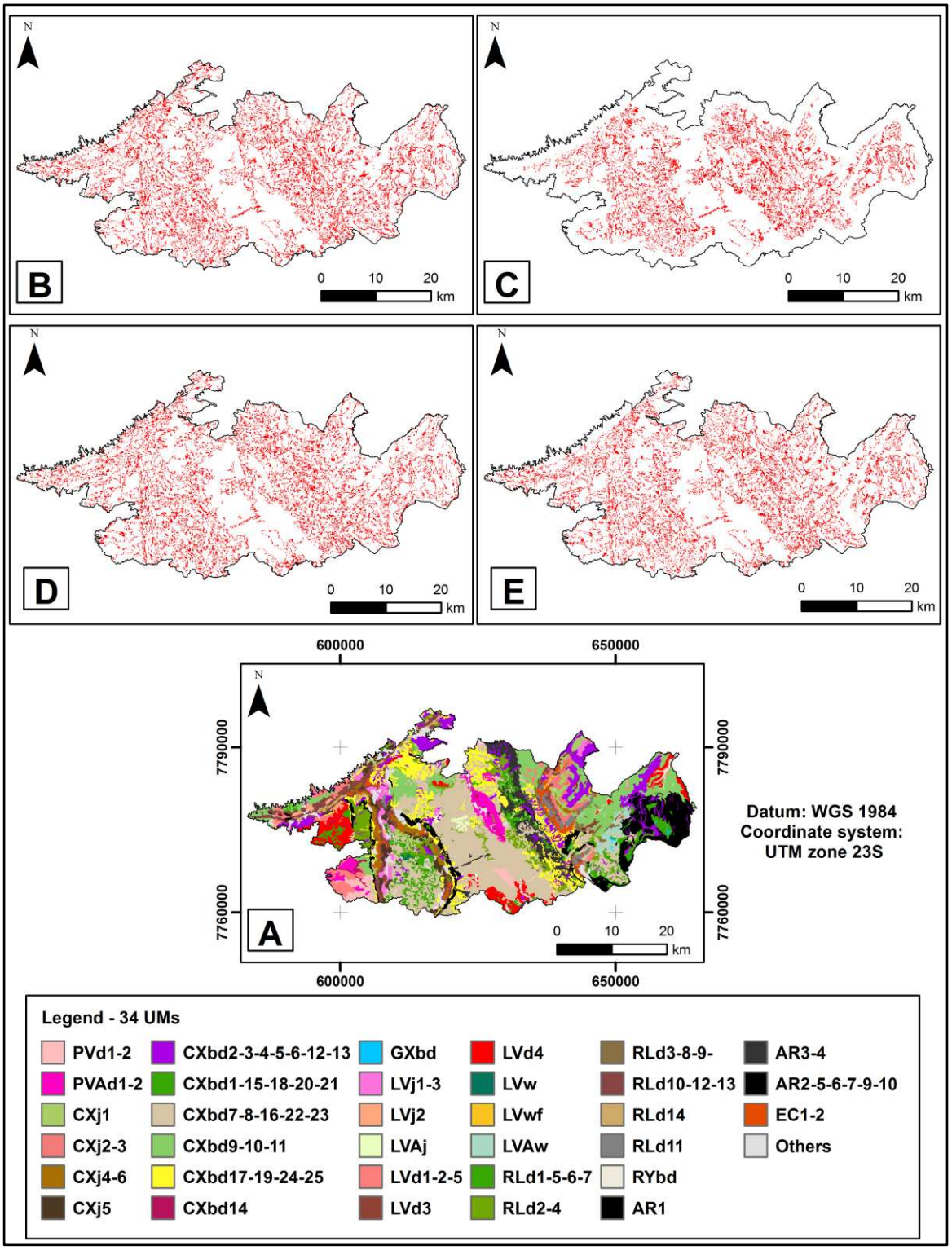


Figura 11. Mapas de incerteza para 34 UMs: discordância entre o mapa de referência e os mapas preditos pelos conjuntos de variáveis selecionados por cada um dos métodos avaliados: A- Mapa de solos de referência (CPRM, 2005); B- Método do Meandecrease Accuracy; C- Método da Correlação; D- Método do Índice Ginni; E- Método do Percentil.

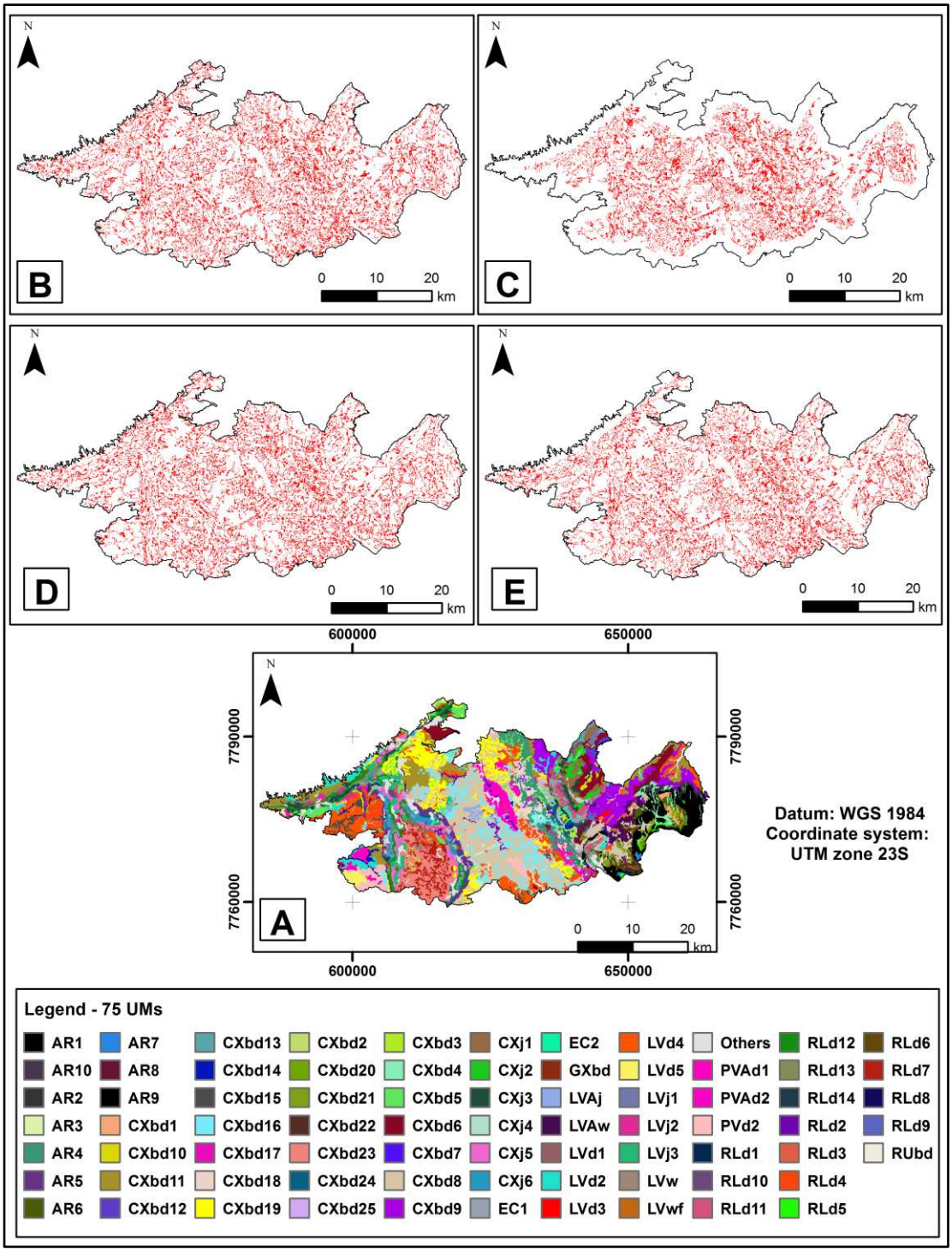


Figura 12. Mapas de incerteza para 75 UMs: discordância entre o mapa de referência e os mapas preditos pelos conjuntos de variáveis selecionados por cada um dos métodos avaliados: A- Mapa de solos de referência (CPRM, 2005); B- Método do Meandecrease Accuracy; C- Método da Correlação; D- Método do Índice Ginni; E- Método do Percentil.

4. DISCUSSÃO

4.1. Métodos de seleção e covariáveis selecionadas

Em todos os métodos de seleção de covariáveis testados observou-se uma diminuição significativa no número de variáveis empregadas, com perdas de acurácia praticamente imperceptíveis. Isto demonstra que existe no conjunto completo de variáveis preditoras significativa redundância de informação e que é possível construir modelos simplificados e eficientes utilizando poucas covariáveis. Houve diferenças significativas no número de variáveis selecionadas no ponto de corte de cada um dos métodos.No entanto, o maior número foi de 34, valor este que representa aproximadamente 25% do valor inicial (141 variáveis). Ainda sim obtiveram-se elevados valores de índice Kappa com cerca de 4 a 18 variáveis (Tabela 5).

Os resultados obtidos nesse trabalho corroboram com aqueles obtidos no estudo feito por Brungar (2014), que também utilizou o ranqueamento feito pelo Random Forest, iniciando com um conjunto total de 113 variáveis e finalizou selecionando um número máximo de 8 variáveis para sua área de estudo. Taghizadeh- Merjardhi et al. (2016), também procederam na seleção de variáveis, no entanto utilizaram o método de seleção Genetic Algorithms (GA), que é inspirado nas forças de seleção natural, e busca os melhores valores para uma função. Esses autores iniciaram os estudos com um conjunto total de 18 variaveis e o reduziram a 46% utilizando este método de seleção de variáveis.

A redução expressiva no número de variáveis encontrada neste trabalho e no trabalho de Brungar (2014), também foi observada por Miller et al. (2015) que partiu de um conjunto inicial de 412 variáveis e obteve subconjuntos selecionados entre 5 e 14 variáveis. Este autor trabalhou com o software de mineração de dados Cubist, demonstrando que, assim como acontece nos classificadores, talvez não exista o melhor método de seleção de variáveis, e sim algumas possibilidades que apresentem resultados similares.

Algumas covariáveis selecionadas, principalmente as categóricas (Mapa de unidades de relevo, Mapa Geomorfológico e Mapa Geológico), destacaram-se por serem selecionadas em praticamente todos os métodos e para todos os agrupamentos de classes de solo. Este fato evidencia a relevância do relevo

morfoestrutural presente na área de estudo, como estratificador da paisagem. Destaca-se a presença das coordenadas geográficas, que podem não ser covariáveis muito explicativas, no entanto demonstraram considerável capacidade preditiva. Estas covariáveis podem ter sido importantes devido a elevada densidade de observações aqui empregada, e ao fato de que existe uma considerável variabilidade ambiental na área principalmente no sentido leste oeste. Todavia podem ser realmente variáveis importantes em algumas áreas de estudo, e talvez possam ser mais empregadas em outros trabalhos, visto que praticamente não são utilizadas como covariáveis preditoras.

Outras duas covariáveis que se mostraram importantes nos agrupamentos selecionados são os mapas temáticos de unidades de relevo e de geomorfologia. Apesar de serem variáveis categóricas vinculadas ao relevo, se mostraram muito importantes no mapeamento de classes de solo na região de estudo. Este fato ressalta a importância do relevo como estratificador dos solos na paisagem estudada. Soma-se ainda a este grupo o mapa temático de geologia que reforça a importância do material de origem no delineamento das UMs em regiões de relevo estrutural com litologias bem definidas.

As demais variáveis selecionadas apresentaram contribuições distintas dos demais fatores de formação de solos (clima, material de origem e organismos). Fato este que evidencia a importância de se utilizar variáveis vinculadas a estes outros fatores de formação de solos, visto que ainda são predominantemente empregadas no mapeamento de solos, as variáveis morfométricas derivadas dos MDEs. Este aspecto se torna ainda mais importante em áreas de relevo mais suavizado, onde os atributos do relevo não são, muitas vezes, os principais estratificadores dos solos na paisagem. Corroboram com esse aspecto, os resultados obtidos por Mosleh et al. (2016) no estudo de mapeamento de propriedades do solo em áreas de planícies no Irã, para o qual tiveram como principais preditores variáveis como NDVI, Perpendicular Vegetation Index (PVI), Carbonate Index (CaI) dentre outras variáveis de relevo como Difuse Radiation, Aspect e Plan curvature.

Os resultados obtidos no presente trabalho apontam que dados geofísicos de gamaespectometria têm grande potencial como covariáveis preditivas em MDS, corroborando com a indicação de Lagacherie (2008). As concentrações dos elementos Th e K, bem como a razão entre ambos, apareceram em 5 dos 13 diferentes conjuntos de variáveis selecionadas

(Tabelas 6, 7 e 8). Além disso, foram selecionadas em todos os agrupamentos de classes, demonstrando seu potencial de predição nos diferentes níveis de detalhamento em termos de número de classes. Ressalta-se também que a concentração de K apareceu na sexta posição de importância (Tabela 7) no melhor conjunto de variáveis selecionado neste trabalho que foi pelo Percentil para 34 UMs.

Conforme Wilford e Minty (2007), o comportamento do K no solo se diferencia do comportamento do U e do Th. De maneira geral o K tem sua concentração decrescida com o avanço do intemperismo. Isto porque este elmento se desprende facilmente dos minerais primários e como íon apresenta elevada solubilidae, tendendo a ser intensamente lixiviado do solo. Porém este pode permanecer no solo associado a minerais de argila como illita ou montmorilonita. E pode estar presente em cristais de muscovita e K-feldspato que ainda estejam inalteradas pelo intemperismo. Em contraste ao comportamento do K, o U e o Th são considerados os produtos mais estáveis em relação ao intemperismo nos perfis de solo. Estes dois elementos apresentam similaridade em seu comportamento geoquímico devido a semelhanças em propriedades químicas como por exemplo o raio iônico. Esta semelhança de raio iônico também é obersvada com o Zircônio (Zr), possibilitando a ocorrência de substituição isomórfica de Zr por Th em minerais de zircão. Além disso, o Th e o U podem se associar a oxihidróxidos de Fe e Al, e estabelecer complexos organominerais com a matéria orgânica do solo. Portanto, é comum encontrar as concentrações mais elevadas de U e Th em solos mais intemperizados.

A relevância dos canais de Th bem como da razão Th/K em relação ao urânio ou da razão U/K, pode ser atribuiída primeiramente a maior estabilidade do Th, conferida pela diferença entre os valores de pH (4,2 e 3,5) que promovem a hidrólise de compostos de U e Th respectivamente, juntamente com a maior concentração média na crosta terrestre do segundo (10,5 ppm) em relação ao primeiro (3 ppm) (BOYLE, 1982).

Os dados de aerogamaespectometria tem sido utilizados em alguns trabalhos de mapeamento de solos, como o de Erbe et al. (2012) que utilizarem estes dados associados aos dados de litologia para tentar distinguir classes de solo na Tailândia. Hermann et al. (2010) fizeram uma quantificação de Th, U e K e correlacionaram os resultados com classes de solos e rochas, também na

Tailândia. No entanto vale ressaltar que estes dados geofísicos ainda têm sido pouco utilizados em trabalhos de MDS e que por apresentarem uma relação direta com a composição geoquímica das rochas e solos, podem ser úteis como variáveis associadas ao material de origem.

As rochas predominantes na região da APA-Sul são naturalmente empobrecidas nas concentrações de ambos os elementos Th e K, visto que em sua maioria são provenientes de sedimentos metamorfisados, e, todavia, a variabilidade de teores em rochas sedimentares ou metasedimentares é muito oscilante. Acredita-se que o potencial preditivo dos dados aerogamaespectométricos possa ser melhor expresso em regiões de rochas com maiores concentrações destes elementos, como por exemplo em regiões do embasamento cristalino onde predominam as rochas graníticas.

4.2. Comparação entre mapas preditos e mapas de referência

Ao se realizar a predição dos mapas com o conjunto de variáveis selecionadas por cada método pode-se identificar o conjunto de variáveis que se mostrou mais eficiente. Embora os valores do índice Kappa tenham sido próximos entre os procedimentos do Percentil, Índice Ginni e Mean Decrease Accuracy, o número de variáveis selecionadas diferiu significativamente. Este fato parece estar relacionado com o tipo de remoção das covariáveis, pois no Percentil as covariáveis eram sempre removidas em grupos, referente ao valor do Percentil, enquanto no índice Ginni e no MDA estas foram removidas uma a uma. A princípio os procedimentos usando o índice Ginni e o Mean Decrease Accuracy selecionaram praticamente a metade do número de variáveis selecionado pelo Percentil. No entanto ao se fazer a comparação entre os mapas preditos e o mapa de referência tornou-se claro que os resultados obtidos pelo método do Percentil foram melhores, pois diferiram estatisticamente dos demais valores de Índice Kappa (Figura 7). Portanto reduzir excessivamente o número de variáveis pode ser prejudicial para o modelo preditivo. No caso deste estudo acredita-se ter alcançado bons resultados com um número muito reduzido de variáveis (4 ou 6), no caso do índice Ginni e do Mean Decrease Accuracy, devido ao número grande de pontos amostrais (32.500).

As Figuras 9, 10, 11 e 12 permitem uma análise visual dos mapas produzidos com cada um dos conjuntos de variáveis selecionadas, e evidenciam que os métodos do Percentil, Índice Gini e Mean Decrease Accuracy proporcionaram a predição de mapas muito semelhantes entre si, sendo muitas vezes impossível distingui-lo visualmente. Somente a partir da comparação estatística entre os valores do Índice Kappa foi possível identificar o melhor mapa gerado que foi predito com o conjunto de variáveis selecionadas pelo método do Percentil para 34 UMs.

4.3. Comparação entre diferentes números de classes preditas

A simplificação das unidades de mapeamento (UMs) do mapa original, que continha 75 UMs foi adotada como uma estratégia para averiguar o comportamento do classificador frente a diferentes níveis de complexidade. A partir dos resultados obtidos nota-se que um nível médio de simplificação foi adequado, visto que os melhores resultados de índice Kappa foram obtidos para os agrupamentos de 24 e 34 UMs. Portanto, tanto a simplificação extrema (9 classes), gerada somente a partir do segundo nível categórico, como as 75 UMs estabelecidas convencionalmente pelo pedólogo na geração do mapa, apresentaram maiores dificuldades para o classificador. No caso das 9 classes, os menores níveis de acurácia se devem ao agrupamento de solos muito distintos dificultando o estabelecimento de um padrão homogêneo. Já no caso das 75 classes o conjunto de variáveis utilizadas não foi capaz de reproduzir todas as nuanças referentes a diversidade de solos mapeados pelo pedólogo no mapa de referência.

5. CONCLUSÕES

– O número de variáveis selecionadas por cada um dos métodos variou significativamente, sendo o maior número atribuído ao método da Correlação, e os menores associados ao Índice Ginni e ao Mean Decrease Accuracy. O método do Percentil selecionou números de variáveis intermediários. No entanto os valores de acurácia (Índice Kappa) tiveram um

comportamento inverso ao número de variáveis, conjuntos de variáveis intermediários e menores apresentaram maiores valores de Índice Kappa.

– O processo de seleção de variáveis se mostrou eficiente em simplificar o conjunto de variáveis, mantendo ótimos valores de acurácia com cerca de 10% do número total de variáveis. Evidenciando que esta é uma etapa que deve ser intrínseca ao procedimento metodológico no MDS, haja visto que a gama de variáveis possível de ser utilizada é enorme e certamente a importância destas irá variar em áreas fisiograficamente distintas.

– A presença de covariáveis geofísicas da gamaespectometria em vários dos conjuntos selecionados demonstra o potencial ainda pouco explorado, em trabalhos de MDS, de uso destes dados como representantes do material de origem.

– O método de seleção de variáveis pela variação do Percentil apresentou-se como método potencial, pois apesar de ter selecionado mais variáveis, apresentou valores de kappa ligeiramente maiores do que o Índice Ginni e o Mean Decrease Accuracy, e forneceu o melhor conjunto de variáveis na comparação pixel a pixel com o mapa de referência.

– O agrupamento de 34 UMs apresentou o melhor valor de Índice Kappa quando predito com o conjunto de variáveis selecionadas pelo método do Percentil, sendo esta a melhor combinação testada neste trabalho. Este resultado evidencia que o nível de detalhamento das classes a serem preditas interfere diretamente na acurácia do mapa gerado.

– A ocorrência significativa de covariáveis vinculados ao clima e ao material de origem no melhor conjunto selecionado (Percentil - 34 UMs), evidenciam a importância de se diversificar as covariáveis preditoras em função dos fatores de formação de solos.

6. REFERÊNCIAS BIBLIOGRÁFICAS

BACCHI OOS, REICHARDT K, OLIVEIRA JCM & NIELSEN DR. Gamma- Ray beam attenuation as an auxiliary technique for the evaluation of the soil water retention curve. Scientia Agricola, Piracicaba, 1998; 55(3): 498–502.

BAGATINI, T.; GIASSON, E.; TESK, R. Seleção de Densidade de Amostragem com Base em Dados de Áreas já Mapeadas para Treinamento de Modelos de Árvore de Decisão no Mapeamento Digital de Solos. Revista Brasileira de Ciência do Solo. 2015; 39:960-967.

BECEGATO VA & FERREIRA FJF. Gamaespectrometria, resistividade elétrica e susceptibilidade magnética de solos agrícolas no noroeste do estado do Paraná. Rev. Bras. Geof., 2005; 23(4): 371–405.

BIERWIRTH, P. Gamma-radiometrics: A remote sensing tool for understanding soils. AGSO. 1996.

BOYLE, R.W. Geochemical prospecting for thorium and uraniuum deposits. Developments in Econommc Geology, n.16. 1982; 71-78p.

BREIMAN, L. Random Forests. Journal of Machine Learning, Kluwer Academic, Netherland. 2001; 45:5-32.

BRUNGARD, C.W.Advancing Digital Soil Mapping and Assessment in Arid Landscapes. Utah State University, All Graduate Theses and Dissertations. Paper 3305. 2014.

CGIAR. Consortium for Spatial Information. SRTM 90m Digital Elevation Data. Available at: <http://srtm.csi.cgiar.org/>. Accessed Fev. 2015.

CONGALTON, R.G.; GREEN, K. **Assessing the accuracy of remotely sensed data: principles and practices**. New York: Lewis Publishers, 1999, 160p.

CPRM. Companhia de Pesquisa de Recursos Minerais. Projeto APA Sul RMBH: estudos do meio físico, pedologia, Shinzato, E., Carvalho Filho, A. - Projeto APA Sul RMBH: Mapa de solos, escala 1:50.000; geologia, mapa geológico, escala 1:50.000 em 3 partes. Sérgio L.da Silva (Org.)., Monteiro, E.A., Baltazar, O.F.; Zucchetti, M.. - geomorfologia, mapa geomorfológico, escala 1:50.000 em 3 partes. Medina, A.I.; Saadi, A. Belo Horizonte: CPRM/EMBRAPA/SEMAD, 2005.

EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA - EMBRAPA. Centro Nacional de Pesquisa de Solo. Procedimentos normativos de levantamentos pedológicos. / Santos, H.G. et al.. Brasília: Embrapa – SPI, 1995. 116p.

ERBE, P.; SCHULEER, U.; STAHR, K.; HERMANN, L. Evaluation of gamma-ray spectrometry aas a tool for soil science. Poster 1.1 Soil care/water and matter International Conference “Sustainable land use and rural development in mountain areas”. Hohenhelm, Stuttgart, Germany. 16-18 April. 2012.

HERMANN, L.; SCHULER, U.; RANGUBPIT, W.; ERBE, P.; SURINKUM, A.; ZAREI, M.; STAHR, K. The potencial of gamaespectrometry to soil mapping. Congress of Soil Science, Soil Solutions for a Changing World. Australia, 2010.

FONNESBECK, B.B. Digital Soil Mapping Using Lands cape Stratification for Arid Rangelands in the Eastern Great Basin, Central Utah. Utah State University, All Graduate Theses and Dissertations. Paper 4525. 2015.

DALMOLIN, R.S.D.; TEM CATEN, A. Mapeamento Digital: Nova abordagem em mapeamento de solos. Investigação Agrária. 2015; 17:77-86.

DIAS, L.M.S. Predição de classes de solo por atributos do meio físico e de sensoriamento remoto em áreas da bacia sedimentar do São Francisco. Dissertação. Mestrado em Agricultura Tropical e Subtropical. Campinas: Instituto Agronômico, 2015. 127p.

GIASSON, E.; ten CATEN, A.; BAGATINI, T.; BONFATTI, B. Instance selection in digital soil mapping: a study case in Rio Grande do Sul, Brazil. Ciência Rural-Santa Maria. 2015.;45:1592-1598.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA-IBGE. Coordenação de Recursos Naturais e Estudos Ambientais. Manual técnico de pedologia. 2ª ed. Rio de Janeiro: IBGE, 2007. 323p.

JENNY, H. Factors of soil formation; a system of quantitative pedology. New York: McGraw-Hill, 1941. 281p.

JIA, S.; LI, H.; WANG, Y.; TONG, R.; LI, Q. Recursive variable selection to update near-infrared spectroscopy model for the determination of soil nitrogen and organic carbon. Geoderma. 2016; p.92-99.

LAGACHERIE, P. Digital Soil Mapping: A State of the art perspectives for Digital Soil Mapping. In: HARTEMINK, A.E.; McBRATNEY, A.B.; MENDONÇA-SANTOS, M.L.; Digital Soil Mapping. Springer, 2008. p.3-14.

LAGACHERIE, P.; MCBRATNEY, A.B., 2007. Spatial soil information systems and spatial soil inference systems: perspectives for Digital Soil Mapping. In: P. Lagacherie, A.B. McBratney and M. Voltz (Eds.), Digital Soil Mapping, an introductory perspective. Developments in soil science, vol. 31. Elsevier, Amsterdam, p.3-24.

LOBATO, L.M.; et al. Projeto Geologia do Quadrilátero Ferrífero - Integração e Correção Cartográfica em SIG com nota explicativa. Belo Horizonte: CODEMIG, 2005. 1 CD-ROM.

MOSLEH, Z.; SALEHI, M. H.; JAFARI, A.; BORUJENI, I. E.; MEHNATKESH, A. The Effectivenes Digital Soil Mapping to Predict Soil Properties Over Low-relief Areas. Environ Monit Assess, february 2016. p.188-195.

McBRATNEY, A.B. et al. On digital soil mapping. Geoderma. 2003;117:3-52.

MENZE, B.H.; KELM, B.M.; MASUCH, R.; HIMMELREICH, U.; BACHERT, P.; PETRICH, W.; HAMPRECHT, F.A. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. Biomed Central LTDA. 2009; p.1-16.

MILLER, B.A.; KOSZINSKI, S.; WEHRHAN, M.; SOMMER, M. Impact of multi-scale predictor selection for modeling soil properties. Geoderma. 2015; p.97-106.

MINTY, B.R.S. Fundamentals of airborne gamma-ray spectrometry. AGSO Journal of Australian Geology e Geophysics, 1997; 17(2), 39-50.

PONTES, M.J.C.; CORTEZ, J.; GALVÃO, R.K.H.; PASQUINI, C.; ARAÚJO, M.C.U.; COELHO, R.M.; CHIBA, M.K.; ABREU, M.F.; MADARI, B.E. Classification of Brazilian soils by using LIBS and variable selection in the wavelet domain. Analytica Chimica Acta. 2009; p.12-18.

SILVA, C.C. Mapeamento digital de classes de solo: aplicação de metodologia na folha Botucatu (SF-22-Z-B-VI-3) e validação de campo. Dissertação. Mestrado em Agricultura Tropical e Subtropical. Campinas: Instituto Agronômico, 2012. 117p.

TAGUIZADEH-MEHRJARDI, R.; NABIOLLAHI, K.; KERRY, R. Digital Mapping of Soil Organic Carbon at Multiple Depths Using Differents Data Mining Techniques in Baneh region Iran. Geoderma. 2016;266:98-110.

ten CATEN, A.; DALMOLIN, R.S.D.; MENDONÇA-SANTOS, M.L.; GIASSON, E. Mapeamento digital de classes de solos: características da abordagem brasileira. Ciência Rural. 2012. 42:1990-1997.

TESK, R.; GIASSON, E.; BAGATINI, T. Produção de um mapa pedológico associando técnicas comuns aos mapeamentos digitais de solo com delineamento manual de unidades de mapeamento. R. Bras Ci do Solo. 2015; 39:950-959.

VAZ CMP, NAIME JM & MACEDO A. Soil particle size fractions determined by gamma-ray attenuation. Soil Science, Baltimore, 1999; 6: 403–410.

VAYSSE, K.; LAGACHERIE, P. Evaluating Digital Soil Mapping aproaches for mapping GlobalSoilMap soil properties from legacy data in Languedoc Roussilon (France). Geoderma Regional. 2015; 4:20-30.

VOHLAND, M.; LUDWIG, M.; THIELE-BRUHN, S.; LUDWIG, B. Determination of soil properties with visible to near- and mid-infrared spectroscopy: Effects of spectral variable selection. Geoderma. 2014; 88-96.

WorldClim – Global Climate Data. Disponível em (<http://www.worldclim.com.br). Acesso em maio de 2015.

WILFORD JR, BIERWIRTH, P.N. & CRAIG, M.A. Application of gamma-ray spectrometry in soil/regolith mapping and geomorphology. AGSO Journal of Australian Geology & Geophysics. 1997; 17- 201–216.
WILFORD, J.; MINTY, B. The use of airborne gamma-ray imagery for mapping soils and understanding landscape processes. Developments in Soil Science, volume 31. 2007.

CONCLUSÃO GERAL

– O desbalanceamento gerado na amostragem, proveniente da desproporcionalidade existente entre o tamanho das áreas das classes é um aspecto que prejudica sensivelmente a acurácia da predição. Buscar estratégias que proporcionem esquemas amostrais mais equalitários é fundamental para que se contemple classes de relevante importância ambiental, mas de pouca expressividade em termos de área. Neste contexto o uso do desbalanceamento, gerado a partir de dados de solo legados, como custo a ser integrado ao método de amostragem cLHS é uma alternativa viável que deve ser avaliada, porém susceptível à qualidade dos dados legados.

– A seleção de variáveis é etapa fundamental ao Mapeamento Digital de Solos, que deve ser implementada como parte intrínseca ao processo metodológico a ser adotado. Além disso, submeter o maior conjunto possível de covariáveis, essencialmete vinculadas a diferentes fatores de formação de solos, pode ser uma maneira imparcial de se obter o melhor conjunto de covariáveis preditivas adequado a cada área de trabalho.

– Os dados de aerogamaespectometria são covariáveis promissoras no Mapeamento Digital de Solos, que podem ser mais utilizadas nos trabalhos desta natureza, visto que são dados de elevado potencial preditivo, e trazem informação de subsuperfície, estando disponíveis em várias regiões do país.

– O Mapeamento Digital de Solos encontra-se em pleno processo de aprimoramento metodológico, e ainda necessita de inúmeros estudos para que possa efetivamente melhorar a informação de solos existente, sendo imprecindível a realização de estudos como este para que se alcance tal condição.