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

MAPEAMENTO DIGITAL DE SOLOS: METODOLOGIAS PARA ATENDER A DEMANDA POR INFORMAÇÃO ESPACIAL EM SOLOS

TESE DE DOUTORADO

Alexandre ten Caten

Santa Maria, RS, Brasil

2011

MAPEAMENTO DIGITAL DE SOLOS: METODOLOGIAS PARA ATENDER A DEMANDA POR INFORMAÇÃO ESPACIAL EM SOLOS

Alexandre ten Caten

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

Orientador: Prof. Dr. Ricardo Simão Diniz Dalmolin

Santa Maria, RS, Brasil

2011

T289m Ten Caten, Alexandre
Mapeamento digital de solos : metodologias para atender a demanda por informação espacial em solos / por Alexandre Ten Caten. – 2011.
106 f. ; il. ; 30 cm

Orientador: Ricardo Simão Diniz Dalmolin
Tese (doutorado) – Universidade Federal de Santa Maria, Centro de Ciências Rurais, Programa de Pós-Graduação em Ciência do Solo, RS, 2011

1. Mapeamento digital de solos 2. Pedometria 3. Levantamento de solos 4. Wavelet 5. Mapa cloroplético 6. Árvore de decisão I. Dalmolin, Ricardo Simão Diniz II. Título.

CDU 631.4:528.7/.9

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

© 2011
Todos os direitos autorais reservados a Alexandre ten Caten. A reprodução de partes ou do todo deste trabalho só poderá ser feita mediante a citação da fonte.
Endereço: Rua Arsênio Machado Soares, 90 ap 204. CEP: 97110-110
Fone (0xx) 55 9945 3935; End. Eletr: acaten@yahoo.com.br

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

A Comissão Examinadora, abaixo assinada,
aprova a Tese de Doutorado

MAPEAMENTO DIGITAL DE SOLOS: METODOLOGIAS PARA ATENDER A DEMANDA POR INFORMAÇÃO ESPACIAL EM SOLOS

elaborada por
Alexandre ten Caten

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

COMISSÃO EXAMINADORA:

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

Fabricio de Araújo Pedron, Dr. (UFSM)

Jean Paolo Gomes Minella, Dr. (UFSM)

Ivan Luiz Zilli Bacic, PhD. (EPAGRI)

Elvio Giasson, Dr. (UFRGS)

Santa Maria, 07 de novembro de 2011.

DEDICATÓRIA

Aos meus pais pela pergunta: ‘- Já fez os tema guri?’.

AGRADECIMENTOS

Às Forças Positivas do Universo que conspiraram para que eu tivesse saúde para concluir essa etapa de minha vida.

À Universidade Federal de Santa Maria por possibilitar o acesso ao conhecimento e as oportunidades de crescimento pessoal em todas as etapas de minha vida acadêmica na instituição.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo financiamento à pesquisa e pela disponibilização do portal de periódicos.

Ao Programa de Pós-Graduação em Ciência do Solo da Universidade Federal de Santa Maria pela oportunidade de cursar o doutorado.

Ao professor Ricardo Simão Diniz Dalmolin pela oportunidade e exemplo de ser humano e profissional.

Aos membros da banca examinadora da defesa de tese de doutorado Ivan Luiz Zilli Bacic, Elvio Giasson, Jean Paolo Gomes Minella e Fabrício de Araújo Pedron pela disponibilidade e contribuições para a melhoria do trabalho.

Aos professores do PPGCS pelos ensinamentos e ao funcionário Heverton Heinz pela prestatividade.

Ao Instituto Federal Farroupilha por viabilizar que eu continuasse meus estudos até a conclusão do doutorado.

Aos colegas e estudantes do Instituto Federal Farroupilha campus Júlio de Castilhos pela compreensão para com minha ausência.

Ao estudante do curso Tecnólogo em Geoprocessamento Luis Fernando Chimelo Ruiz pela valorosa ajuda com a produção dos dados.

Aos colegas do setor de Pedologia e do PPGCS pelas inúmeras trocas.

À Leo por ter estado ao meu lado durante todos esses anos de minha formação acadêmica.

Obrigado

EPÍGRAFE

But there is still *terra incognita* ahead of soil science ...
(Sabine Grundwald)

RESUMO

Tese de Doutorado
Programa de Pós-Graduação em Ciência do Solo
Universidade Federal de Santa Maria

MAPEAMENTO DIGITAL DE SOLOS: METODOLOGIAS PARA ATENDER A DEMANDA POR INFORMAÇÃO ESPACIAL EM SOLOS
AUTOR: ALEXANDRE TEN CATEN
ORIENTADOR: RICARDO SIMÃO DINIZ DALMOLIN
Data e Local da Defesa: Santa Maria, 07 de novembro de 2011.

O solo é cada vez mais reconhecido como tendo um importante papel nos ecossistemas, assim como para a produção de alimentos e regulação do clima global. Por esse motivo, a demanda por informações relevantes e atualizadas em solos está em uma crescente. O Mapeamento Digital de Solos (MDS) possibilita gerar essas informações demandadas em diferentes resoluções espaciais e com indicadores de qualidade associados. O objetivo deste estudo foi analisar as principais abordagens metodológicas utilizadas nos mapeamentos digitais de classes de solos através de uma revisão de literatura dos trabalhos nacionais, assim como propor procedimentos para a análise dos dados a serem utilizados em projetos de mapeamento digital de classes de solos. O emprego de técnicas de MDS para o mapeamento de classes de solos é recente no país, a primeira publicação nesse sentido ocorreu apenas em 2006. Entre as funções preditivas utilizadas predomina o emprego da técnica de regressões logísticas. Quanto à avaliação da qualidade dos modelos preditivos o emprego da matriz de erros e do índice kappa têm sido os procedimentos mais usuais. O emprego da transformada wavelet mostrou-se como uma metodologia de grande potencial para a análise da resolução espacial de máxima variabilidade de atributos de terreno a serem usados em projetos de MDS. A metodologia proposta de exclusão dos dados oriundos de covariáveis ambientais localizadas na bordas dos polígonos de solos possibilitou a geração de modelos por Árvore de Decisão (AD) menos complexos e mais precisos. Assim como o volume de dados necessários para o treinamento de modelos preditivos por AD está entre cinco e 15% do conjunto total de dados como mostrou este estudo. Observações coletadas a campo indicaram uma acurácia dos mapas preditos próxima a 70% para os modelos oriundos dessas densidades de amostragem.

Palavras-chave: ondaleta, árvore de decisão, pedometria, levantamento de solos.

ABSTRACT

Doctoral Thesis
Graduation Program in Soil Science
Federal University of Santa Maria

DIGITAL SOIL MAPPING: METHODS TO MEET THE DEMAND FOR SOIL SPATIAL INFORMATION

AUTHOR: ALEXANDRE TEN CATEN
ADVISER: RICARDO SIMÃO DINIZ DALMOLIN
Defense Place and Date: Santa Maria, November 07th, 2011.

Soil has increasingly being recognized as having an important role in ecosystems as well as for food production and global climate regulation. For this reason, the demand for relevant and updated information on soil is increasing. Digital Soil Mapping (DSM) provides this information at different spatial resolution with associated quality indicators. The aim of this study was to analyze the main methodological approaches used for DSM of soil classes through a literature review of national researches and to propose procedures for data analysis in DSM projects of soil classes. The use of DSM techniques for mapping soil classes in Brazil is recent, the first publication on this subject occurred only in 2006. Among the predictive functions, logistic regressions is the predominantly used technique. Quality evaluation of the predictive models employed error matrix and kappa index in most cases. The use of wavelet transform proved to be a methodology of great potential for analyzing the spatial resolution of terrain attributes maximum variability. The proposed methodology of data exclusion for environmental covariates located too near at the border of soil classes polygons has enabled the generation of less complex and more accurate Decision Tree (DT) models. It was also shown that the amount of data required for DT model training is between five and 15% of the total data set. Collected field observations indicated a predicted accuracy close to 70% for DT models produced by those sampling densities.

Key words: wavelet, decision tree, pedometric, soil survey

SUMÁRIO

INTRODUÇÃO

O Mapeamento Digital de Solos (MDS), em nível nacional, está ultrapassando a posição de nova possibilidade para pesquisa e caminha para sua consolidação como técnica de auxílio ao mapeamento convencional de solo. Esse último há muito tido e dito como insuficiente para fornecer as informações de forma quantitativa e qualitativa de acordo com as demandas.

Algumas iniciativas em termos nacionais têm demonstrado a maior atenção dos cientistas brasileiros a esta linha de pesquisa. Entre elas podem ser citadas a criação da Comissão Especializada em Pedometria dentro da divisão Solo no Espaço e no Tempo pela Sociedade Brasileira de Ciência do Solo (SBCS); a formação da RedeMDS, sob os auspícios da EMBRAPA Solos, visando à reunião de pesquisadores interessados em desenvolver essa área do conhecimento no país; além do aumento das pesquisas em MDS realizadas no Brasil, que tem se refletido em um maior número de publicações científicas relacionadas a esse tema.

Este estudo parte do pressuposto de que o MDS já é também uma realidade nacional e, como já foram realizadas as primeiras demonstrações em dissertações e teses, faz-se necessário agora instrumentalizar os pesquisadores de procedimentos metodológicos para a aplicação dessa técnica de mapeamento nas situações em que a informação espacial em solos seja necessária. Esta tese de doutorado está dividida em quatro capítulos, os quais representam indagações metodológicas dos autores em projetos de MDS.

O objetivo geral deste estudo foi avaliar procedimentos metodológicos a serem empregados em projetos de mapeamento digital de classes de solos. Esse objetivo geral foi apoiado em uma revisão de literatura dos trabalhos nacionais relacionados ao mapeamento digital de classes de solos. A revisão de trabalhos nacionais teve como intuito identificar métodos os quais necessitem de uma maior contribuição por parte da pesquisa em MDS.

Os objetivos específicos do presente trabalho foram: (i) avaliar o potencial da transformada wavelet no estudo da variabilidade espacial e na definição da resolução espacial de atributos de terreno potencialmente empregáveis ao MDS; (ii) testar um método de exclusão dos dados de covariáveis preditoras localizados nos locais de maior variabilidade na transição entre os polígonos de classes de solos; (iii) verificar a influência do volume de dados da área de treinamento utilizado para gerar os modelos preditivos por Árvore de Decisão (AD) aplicados ao MDS.

O primeiro capítulo traz a revisão de literatura dos resultados de pesquisa realizados por pesquisadores brasileiros no mapeamento digital de classes de solos. A importância que esse tema vem ganhando entre os pesquisadores se reflete no número de trabalhos produzidos. São cada vez mais frequentes as apresentações na forma de pôsteres e orais nos eventos nacionais relacionados à ciência do solo, assim como no número de publicações em periódicos científicos. Até onde se pode constatar, uma revisão de literatura com esse enfoque é inédita no país, o que demonstra a importância de se reunir e discutir as principais características dos trabalhos já realizados. Essa revisão irá possibilitar uma perspectiva mais ampla dos caminhos que o mapeamento automatizado de solos vem seguindo no país, além de nortear trabalhos e demandas futuras.

Na sequência da tese, são apresentados os resultados de um estudo no qual se buscou aplicar a análise Wavelet (português Ondaleta) para explorar a variabilidade espacial de atributos de terreno. Entre as covariáveis preditoras mais aplicadas no MDS estão os atributos de terreno ligados ao fator relevo de formação do solo. A geração de alguns desses atributos de terreno é sensível à resolução espacial em que se está trabalhando. Esse é o caso, por exemplo, do índice de posição topográfica que é utilizado para classificar a paisagem em classes morfológicas. Neste capítulo da tese, foi aplicada a análise Wavelet nos atributos de terreno elevação, declividade, curvatura em perfil e índice de umidade topográfica, visando definir a resolução espacial mais apropriada para representar a variabilidade de cada um dos quatro atributos na área do estudo, além de demonstrar a aplicação dessa forma de análise espacial de dados. A aplicação dessa técnica para o estudo de dados em ciência do solo no Brasil ainda é incipiente como se pode constatar na fase de revisão bibliográfica do segundo capítulo.

No terceiro capítulo, são apresentados os resultados de uma estratégia de pré-processamento dos dados a serem utilizados no mapeamento digital de classes de solos. Os mapas de classes de solos existentes são comumente utilizados para treinar os modelos preditivos. Nos mapas de solos, o delineamento das unidades de mapeamento é feito através da interpretação visual de pares estereoscópicos. Intrínseco a esse método, está presente a subjetividade da posição mais adequada para o limite entre os polígonos de solos. A utilização desses polígonos como referência para o treinamento de modelos preditivos implicará a adição de informações desviadas. O estudo demonstra o efeito da utilização de diferentes conjuntos de dados na qualidade de mapas de classes de solos, com e sem pré-tratamento. O pré-tratamento proposto é a utilização da função *buffer* para direção interna aos polígonos

originais. Com isso, informações próximas às bordas dos polígonos não são utilizadas para treinar os modelos preditivos por AD.

No quarto e último capítulo, foi avaliado o efeito do volume de dados utilizado para gerar os modelos preditivos. A disponibilização de modelos digitais de terreno e imagens com resoluções espaciais cada vez maiores faz com que um grande volume de dados esteja disponível e tenha de ser manipulado no MDS. O intuito deste estudo foi avaliar, dentro de um universo de informações disponíveis, qual o volume de dados mais adequado para a geração de modelos preditivos por AD a serem aplicados no mapeamento de classes de solos, visando evitar a manipulação desnecessária de dados, assim como evitar que um conjunto muito pequeno de dados implique a impossibilidade dos modelos preditivos em capturar a complexidade inerente à distribuição espacial de classes de solos.

ARTIGO 1- Mapeamento digital de classes de solos: características da abordagem brasileira

Digital soil mapping: characteristics of the brazilian approach

-REVISÃO BIBLIOGRÁFICA-**

RESUMO

O solo é cada vez mais reconhecido como tendo um importante papel nos ecossistemas, assim como para a produção de alimentos e regulação do clima global. Por esse motivo, a demanda por informações relevantes e atualizadas em solos está em uma crescente. Pesquisadores em ciência do solo estão sendo demandados a gerar informações em diferentes resoluções espaciais e com qualidade associada dentro do que está sendo chamado de Mapeamento Digital de Solos (MDS). Devido ao crescente número de trabalhos relacionados ao MDS, faz-se necessário reunir e discutir as principais características dos estudos relacionados ao mapeamento automatizado de classes de solos no Brasil, o que irá possibilitar uma perspectiva mais ampla dos caminhos, além de nortear trabalhos e demandas futuras. O mapeamento de classes de solos empregando técnicas de MDS é recente no país, a primeira publicação nesse sentido ocorreu apenas em 2006. Entre as funções preditivas utilizadas, predomina o emprego da técnica de regressões logísticas. O fator de formação relevo foi empregado na totalidade dos estudos revisados. Quanto à avaliação da qualidade dos modelos preditivos, constatou-se que a matriz de erros e o índice kappa têm sido os procedimentos mais usuais. A consolidação dessa abordagem automatizada como ferramenta auxiliar ao mapeamento convencional passa pelo treinamento dos jovens pedólogos para a utilização de tecnologias da geoinformação e de ferramentas quantitativas dos aspectos de variabilidade do solo.

Palavras-chave: pedometria, classes de solos, levantamento de solos.

ABSTRACT

Soil is increasingly being recognized as having an important role in ecosystems, as well as for food production and global climate regulation. For this reason, the demand for relevant and updated soil information is increasing. Soil science researchers are being demanded to produce information in different spatial resolutions with associated quality in what is being called Digital Soil Mapping (DSM). Due to an increasing number of papers related to the DSM in Brazil, it is necessary to discuss the main characteristics of those studies related to the automated mapping of soil classes, which will enable a broader perspective of the subject and guide future works and demands. The mapping of soil classes using DSM techniques is recent in the country, the first publication in this topic occurred just in 2006. Among the predictive functions the predominant is logistic regression. The soil formation factor relief was used in all studies reviewed. Quality of predictive models was evaluated employing error matrix and kappa which were the most common procedures. The consolidation of this automated approach as an auxiliary tool to the conventional soil mapping will demand training of young soil scientists to use geoinformation technologies and quantitative tools to handle aspects of soil variability.

Key words: pedometric, soil classes, soil survey.

INTRODUÇÃO

Demandas para a ciência do solo

Para alimentar a população mundial, em 2050, mais de um bilhão de hectares nativos terão de ser convertidos em áreas agricultáveis, caso os modos de produção continuem como os atuais. Visando tornar a agricultura sustentável, é preciso monitorar os efeitos de sua prática em todo o planeta. Historicamente, estratégias agrícolas têm focado em lucratividade e produção. É necessário que esta estratégia seja convertida em sustentabilidade ambiental, segurança alimentar, saúde humana e no bem estar econômico e social. Para que isso seja

alcançado, um conjunto de dados, em diferentes escalas, precisa ser coletado e disponibilizado aos agricultores, pesquisadores e políticos (Sachs et al., 2010).

O solo é cada vez mais reconhecido como tendo um importante papel não só para os ecossistemas, como também para a produção de alimentos e regulação do clima global. Por esse motivo, a busca por informações relevantes e atualizadas em solos está em uma crescente. Contudo, a comunicação da informação acerca do solo é um desafio devido à divergência de termos, desatualização, generalização e imprecisão dos métodos (Sanchez et al., 2009). A pesquisa em solo tem respondido a suas próprias dúvidas e não tem disponibilizado respostas quantitativas a antigas perguntas, de maneira a ser diretamente utilizada por usuários do solo e tomadores de decisão (Hartemink & McBratney, 2008).

Para Hartemink & McBratney (2008), o conhecimento e a pesquisa em solos vêm sendo mais valorizados na medida em que será necessário alimentar 8 bilhões de habitantes, produzir energia a partir da agricultura e aumentar a produção animal embora, mundialmente, a informação sobre o solo não seja acurada ou digitalmente disponível nem atualizada. Entre as oportunidades para a ciência do solo, estão as várias técnicas e métodos já disponíveis aos cientistas de solos, além dos softwares e hardwares que necessitam serem explorados e melhorados. Uma das oportunidades é o Mapeamento Digital de Solos (MDS). No MDS, amostras de solos e covariáveis ambientais são coletadas por metodologias tradicionais ao nível do solo, em aviões ou por satélites. Esses dados são empregados em modelos preditivos que possibilitam gerar novas informações com estatísticas de qualidade associadas (Minasny et al., 2008).

Mapeamento Digital de Solos (MDS)

O termo Pedometria foi criado pelo pesquisador Alex McBratney da Universidade de Sidnei para descrever o estudo quantitativo da variação do solo (Burrough et al., 1994). A predição e o Mapeamento Digital de Solos (MDS) – *Digital Soil Mapping* – tiveram suas

bases estabelecidas por McBratney et al. (2003) e definidas por Lagacherie & McBratney (2007) como “a criação e a população de sistemas de informação espacial de solos por meio de modelos numéricos visando inferir as variações espaciais e temporais de classes e propriedades do solo, a partir de observações, conhecimento e dados de covariáveis ambientais relacionados”.

Como principal aplicação dessa abordagem está a predição, por meio de modelos matemáticos, das classes e propriedades de solos e o mapeamento digital dos resultados de forma contínua e espacial, criando a possibilidade de organizar um amplo conjunto de dados para análise e interpretações em qualquer época, não sendo o mapa o único produto (McBratney et al., 2003). Essa abordagem iniciou nos anos 70 e teve grande desenvolvimento nos anos 80 devido aos avanços tecnológicos nas áreas de tecnologia da informação, sensoriamento remoto, estatística, modelagem, posicionamento global, sistemas de medida e, mais recentemente, acesso instantâneo à informação através da rede mundial de computadores.

Para Sanchez et al. (2009), entre as principais etapas do MDS estão: (i) reunião e unificação das informações disponíveis, representando os fatores de formação como modelos digitais de elevação, mapas de vegetação, clima e geológicos, além de informações disponíveis em relatórios e mapas de solos sobre sua distribuição espacial; (ii) mapeamento de classes e propriedades do solo através da aplicação de relações matemáticas entre essas e os fatores de formação do solo; (iii) aplicação das informações mapeadas com vistas a gerar novas informações, as quais são mais difíceis de serem medidas do que estimadas; (iv) a informação gerada na etapa anterior é confrontada com as atividades antrópicas, gerando mapas de conflitos, riscos e possibilidades e, por fim, (v) a geração de um conjunto de medidas e recomendações baseado no conhecimento da distribuição espacial do solo.

A presente revisão bibliográfica justifica-se na medida em que é crescente o número de trabalhos publicados a respeito da aplicação da técnica de mapeamento digital de classes de solos no Brasil. Faz-se necessário reunir e discutir as principais características dos trabalhos já realizados, o que irá possibilitar uma perspectiva mais ampla dos caminhos que o mapeamento automatizado de solos vem seguido no país, além de nortear trabalhos e demandas futuras.

DESENVOLVIMENTO

Metodologia adotada

Artigos relacionados ao mapeamento digital de classes de solos realizados no Brasil foram reunidos neste estudo. Para que o artigo publicado fosse selecionado para esta revisão, deveria ter aplicado o MDS em um sentido *stricto sensu*. Interpreta-se o *stricto sensu* como o emprego de covariáveis preditoras de uma ou mais funções matemáticas e de um conjunto de dados de treinamento. Abordagens *lato sensu* do MDS, em que não ocorre uma classificação numérica de solos, mas apenas uma delimitação de classes de solos apoiada por sistemas informatizados, como realizado por Sarmento et al. (2008), não foram incluídas neste estudo.

Para este trabalho, efetuou-se uma busca nas plataformas de dados *Scielo*, *Scopus* e *Web of Science*, utilizando palavras-chave relacionadas ao tema 'mapeamento digital de classes de solos'. A busca não fez distinção entre revistas científicas nacionais ou estrangeiras.

O MDS no Brasil é recente

Seguindo os critérios apresentados, dez estudos foram reunidos e as principais informações foram tabuladas para fomentar a discussão nesta revisão. O mapeamento de classes de solos empregando técnicas de MDS é recente no país (Tabela 1). A primeira publicação nesse sentido ocorreu apenas em 2006 (Giasson et al., 2006), mesmo ano em que foi sediado no Rio de Janeiro o 2° Workshop Global em Mapeamento Digital de Solos

126 (Hartemink et al., 2008). Esse evento pioneiro visou divulgar entre os cientistas brasileiros
127 que estudam solo os métodos e técnicas empregados no MDS.

128 **Tabela 1** – Informações descritivas de dez estudos em mapeamento digital de classes de solos
129 realizados no Brasil.

Estudos	Modelos	Fatores	Atributos usados	Localização	Escala	Área (km²)
Giasson et al.(2011)	AD	r, s	dec , pla, per, umi, cur, dir, acu, com, pod	Bento Gonçalves (RS)	1:10.000	6,14
ten Caten et al. (2011a)	REG	r, c	ele, dec , pla, per, dis, con, umi, sed, rad	São Pedro do Sul (RS)	1:50.000	874
ten Caten et al. (2011b)	REG	r, c	ele, dec , pla, per, dis, con, umi, sed, rad	São Pedro do Sul (RS)	1:50.000	874
ten Caten et al. (2011c)	REG	r, c	ele, dec, pla, per, dis, con, umi, sed, rad	São Pedro do Sul (RS)	1:50.000	874
Chagas et al.(2010)	NEU	r, s, p, o	ele, dec, pla, umi, asp, geo, arg, oxi, ndv	Bacia hidrográfica do Rio São Domingos (RJ)	1:100.000 a 1:50.000	9,9
Coelho & Giasson (2010)	REG, AD, BAY, CAR, LMT	r, s	dec, pla, per, umi, cur, dir, acu	Ijuí, Bozano e Coronel Barros (RS)	1:50.000	1.018
Crivelenti et al. (2009)	AD	r, p	dec, pla, per, dis, con, geo	Carta topográfica Dois Corregos (SP)	1:100.000	772
Carvalho et al.(2009)	NEB	r, p, o	ele, dec, geo, veg	Mucugê (Ba)	1:50.000	382,2
Figueiredo & Giasson (2008)	REG	r, s	dec, pla, per, dis, umi, dir, acu, com	Ibirubá e Quinze de Novembro (RS)	1:80.000	720
Giasson et al. (2006)	REG	r, s	dec, pla, per, umi, cur, dir, acu, com, pod	Sentinela do Sul (RS)	1:50.000	253

130

131 REG: regressões logísticas múltiplas, AD: árvore de decisão, BAY: classificação Bayes, CAR: classificação
132 hierárquica CART, LMT: classificação hierárquica LMT, NEU: redes neurais, NEB: lógica nebulosa, r: relevo,
133 c: clima, s: solo, p: material de origem, o: organismos, ele: elevação, dec: declividade, pla: curvatura planar, per:
134 curvatura de perfil, dis: distância à drenagem, con: área de contribuição, umi: índice de umidade topográfica,
135 sed: capacidade de transporte de sedimento, cur: curvaturas combinadas, dir: direção de fluxo, acu: fluxo
136 acumulado, com: comprimento de fluxo, pod: índice de poder de córrego, asp: aspecto, geo: geologia, arg: índice
137 de argila, oxi: óxido de ferro, ndv: NDVI, veg: vegetação, rad: radiação relativa disponível.
138

Esse despertar mais recente dos cientistas brasileiros para o MDS pode estar relacionado (i) à disponibilização bem mais tardia de software e hardware aqui no país; (ii) ao conservadorismo de muitos pedólogos que relutam em utilizar sistemas automatizados capazes de contribuir para o mapeamento de solos; (iii) à carência de pessoal qualificado para o emprego da tecnologia da informação na ciência do solo; (iv) à popularização mais recente no Brasil de tecnologias, como sistema de posicionamento global e sensoriamento remoto, que possibilitam gerar dados sobre o ambiente visando a geração dos modelos.

Por outro lado, o mapeamento de classes de solos empregando técnicas automatizadas já vem sendo praticado há muitos anos em outros países. Webster & Burrough (1972) demonstraram, há quase quatro décadas, o emprego de estatística multivariada aplicada em sistemas informatizados para mapear classes de solos no sul da Inglaterra.

Estudos relacionados ao MDS estão se tornando mais frequentes. Em 2011, foram quatro trabalhos publicados (Giasson et al., 2011; ten Caten et al., 2011a; ten Caten et al., 2011b; ten Caten et al., 2011c). Com a criação da Comissão Especializada em Pedometria dentro da divisão Solo no Espaço e no Tempo pela Sociedade Brasileira de Ciência do Solo (SBCS), haverá uma maior divulgação do MDS entre cientistas do solo no Brasil.

Modelos preditivos empregados

Entre as funções matemáticas utilizadas, predomina o emprego da técnica de regressões logísticas (Tabela 1). Através de uma relação linearizada entre covariáveis preditoras e classes de solos, as regressões logísticas geram um valor de pertinência para cada classe de solos a ser mapeada sobre a paisagem (Giasson et al., 2006). Esse valor de probabilidade de ocorrência é posteriormente avaliado entre todas as classes de solos para um mesmo ponto da paisagem, sendo que entre elas a de maior valor é atribuída à coordenada em análise, gerando-se um mapa de classe de solos. As regressões logísticas apresentam como desvantagem uma sensibilidade à proporção relativa entre classes de solos no conjunto de

164 treinamento dos modelos (ten Caten et al., 2011a) e uma relativa complexidade para a
165 interpretação dos parâmetros estatísticos dos modelos logísticos gerados (Kempen et al.,
166 2009).

167 Metodologias que busquem relacionar fatores de formação e classes de solos de forma
168 mais similar ao raciocínio do pedólogo têm sido aplicadas pelos estudos que empregam a
169 técnica de mineração de dados por Árvore de Decisão (AD) (Crivelenti et al., 2009; Giasson
170 et al., 2011). O emprego de AD é fruto de pesquisas que buscam aplicar técnicas robustas
171 para a extração de padrões em grandes conjuntos de dados (Witten & Frank, 2005). A
172 abordagem por AD tem sido empregada por apresentar a vantagem de possibilitar a expressão
173 das relações solo-paisagem de maneira explícita (Kheir et al., 2010a). Além de permitir o
174 agrupamento e a busca por padrões, a AD possibilita o entendimento de como esses dados são
175 inter-relacionados (Kheir et al., 2010b). Nos estudos até aqui realizados no país, não se tem
176 dado uma ênfase à análise das regras de decisão geradas durante a modelagem. Isso permite
177 afirmar que a AD vem sendo empregada mais pela sua robustez como técnica preditiva, com
178 mais ênfase ao mapa final gerado, do que ao seu potencial em explicitar e esclarecer as
179 relações entre fatores de formação e classes de solos.

180 Entre as estratégias referidas como 'caixa preta' (Ballabio, 2009), a técnica de redes
181 neurais artificiais (RNA) foi empregada por Chagas et al. (2010). O uso de redes neurais
182 possibilita uma grande acurácia na predição, contudo, os problemas aparecem quando a rede
183 precisa ser aplicada como conhecimento formalizado devido à complexidade da rede e dos
184 pesos empregados (Qi & Zhu, 2003). Kheir et al. (2010b) concordam com essa dificuldade
185 para a implementação das RNA. Para esses autores, as redes neurais não possibilitam um
186 modelo de fácil compreensão que permita os pesquisadores ter um entendimento completo da
187 natureza dos dados que estão sendo analisados. Também entre as estratégias referidas como

188 modelo ‘caixa preta’ está o emprego de Máquinas de Vetores de Suporte (Ballabio, 2009), o
189 qual ainda não foi empregado em nenhum estudo no Brasil.

190 **Fatores de formação empregados**

191 O fator de formação relevo foi empregado na totalidade dos estudos revisados (Tabela
192 1). Possivelmente isso se deva a sua ampla disponibilidade através de Modelo Digital de
193 Elevação (MDE) oriundo do SRTM (*Shuttle Radar Topography Mission*) e de cartas
194 topográficas. Também contribui para sua aplicação a possibilidade de derivar a partir desse
195 um maior número de covariáveis preditoras; a clara relação existente entre o relevo e o padrão
196 de distribuição espacial das classes de solos, além da resolução espacial dos arquivos raster
197 associados a este fator.

198 Uma das principais fontes de informação para o MDS é o MDE. MDE derivados de
199 cartas topográficas e das plataformas SRTM e ASTER necessitam ser avaliados, bem como
200 novas fontes de dados de maior resolução espacial, a exemplo da plataforma TerraSAR-X, os
201 quais carecem de investigação sobre sua aplicabilidade. O posicionamento por satélite, que
202 tem sido realizado a partir da constelação americana (GPS) e russa (GLONASS), apresentará
203 ganhos de qualidade no posicionamento com a disponibilização dos sistemas europeu
204 (Galileo) e chinês (Compass) em um futuro breve.

205 Os fatores material de origem e clima foram utilizados com menor frequência (Tabela
206 1). Isso decorre possivelmente da generalização dos mapas disponíveis dessas informações.
207 Os mapas geológicos amplamente disponíveis para todo o território nacional encontram-se na
208 escala 1:1.000.000 (CPRM, 2011). Já a informação acerca do clima, além da questão da falta
209 de resolução nos mapas disponíveis, existe ainda o fato de que os solos tenham se formado
210 em tempos pretéritos, podendo haver uma baixa correlação entre condições climáticas atuais e
211 as classes de solos.

O Brasil apresenta desafios bem particulares ao mapeamento digital de classes de solos. Entre os desafios está o uso de técnicas de sensoriamento remoto para gerar informações acerca dos fatores de formação do solo. Mapas, como de uso da terra e índices de vegetação (NDVI), apresentam grandes desafios à investigação devido à ação antrópica que modifica a cobertura natural e à presença de nuvens frequentes em várias regiões desse país predominantemente tropical. Isso ocorre também no uso do sinal refletido pelo solo para os sensores orbitais, que acaba sendo comprometido pela cobertura do solo em áreas agrícolas onde se pratica o plantio direto e/ou pela cobertura vegetal natural em muitas regiões do país.

Distribuição e dimensão dos estudos

Entre os dez estudos analisados, sete foram realizados no Estado do Rio Grande do Sul (Tabela 1). Isso pode ser atribuído à presença de dois grupos de pesquisa que desenvolvem atividades nessa linha, naquele Estado. É importante que essa técnica seja divulgada e avaliada em outras regiões do país. Nesse aspecto, a Comissão Especializada em Pedometria da divisão Solo no Espaço e no Tempo (SBCS) terá um papel muito importante para a consolidação das pesquisas em MDS, em todas as regiões do país. Essa informação contrasta com os vazios de levantamentos de solos reportados por Mendonça-Santos & Santos (2007). Esses autores descrevem que as regiões norte e noroeste, sobretudo grandes áreas de floresta Amazônica, possuem informação apenas em pequena escala sobre os solos da região. Ainda conforme esses autores, isso decorre da dificuldade de acesso, políticas de preservação da floresta e a baixa fertilidade do solo, que têm prejudicado os levantamentos de solos na região.

Como a aplicação do MDS implica em uma fase de treinamento dos modelos preditivos, o efetivo auxílio dessa técnica para o aumento do conhecimento a respeito da distribuição espacial dos solos na região norte e noroeste só irá ocorrer a partir da disponibilização dessas informações. Esse é um desafio a ser superado. Coletas a campo e por

237 técnicas não proximais deverão gerar um grande volume de dados, caso se deseje um
238 mapeamento em escalas mais detalhadas.

239 O MDS aplicado à predição de classes de solos no Brasil tem ocorrido
240 predominantemente em nível semidetalhado (1:50.000) (Tabela 1), possivelmente devido ao
241 emprego de Cartas Topográficas para a geração de MDE a fim de produzir atributos de
242 terreno utilizados para treinar os modelos preditivos. Embora uma das prerrogativas à
243 aplicação de técnicas de MDS seja possibilitar o mapeamento a partir de demandas
244 específicas pelo conhecimento da distribuição do solo (Carré et al., 2007), isso não tem sido
245 reportado nas publicações nacionais. Dessa forma, o volume de informações e,
246 consequentemente, a escala do mapeamento têm sido definidos por outros critérios que não a
247 demanda pela informação.

248 Mapas de solos, disponíveis para o limite de alguns municípios, têm sido o conjunto
249 de dados mais utilizado para treinamento dos modelos preditivos (Tabela 1). Estudos que
250 utilizam os limites administrativos dos municípios podem ter sua aplicação limitada pelo fato
251 de uma Bacia Hidrográfica (BH) estar distribuída em um conjunto de municípios. Esse
252 mesmo problema ocorre caso seja utilizado dados oriundos de cartas topográficas. Esse fator
253 tem condicionado a extensão territorial da maioria dos trabalhos. Apenas o estudo
254 desenvolvido por Chagas et al.(2010) foi realizado a partir das dimensões de uma BH. As BH
255 têm sido indicadas como unidade básica para questões importantes. No Brasil, a Lei Federal
256 nº9.433/97 estabelece a bacia hidrográfica como unidade territorial para aplicação da Política
257 Nacional de Recursos Hídricos. Possivelmente seria a unidade mais adequada para novos
258 mapeamentos pela técnica de MDS, uma vez que é crescente a demanda por informação nessa
259 unidade, disponibilizando informação para estudos futuros.

Procedimentos metodológicos

Os estudos até aqui realizados utilizaram de diferentes densidades de amostragem (Tabela 2). Como os estudos são realizados a partir de dados no formato matriz (raster), a unidade mínima desse modelo de representação dos dados, o pixel, é utilizado como amostra para ajuste dos modelos preditivos. O conjunto de dados utilizados para treinar os modelos representou desde menos de um por cento (Giasson et al., 2011) até a totalidade das amostras (Carvalho et al., 2009), indicando que existe uma carência pela padronização do número de amostras a ser utilizada na fase de treinamento dos modelos. Estratégias apropriadas para a seleção das observações conduzem a melhores resultados do que quando os modelos são ajustados ao total de observações disponíveis (Schmidt et al.; 2008).

A totalidade dos estudos utilizou de MDE, como já mencionado, possivelmente devido a questões ligadas ao fator de formação relevo. Entre as fontes de MDE utilizadas, destaca-se o SRTM (*Shuttle Radar Topography Mission*). Esse dado tem sido utilizado em sua resolução original de 90 m ou no formato Topodata (Valeriano & Rossetti, 2010), que, por sua vez, consiste em uma interpolação dos dados SRTM para 30 m. As curvas de nível presentes nas cartas topográficas também têm sido utilizadas como fontes de dados altimétricos. A utilização de tecnologias como o LIDAR, Scanners 3D e GNSS RTK para a obtenção da informação sobre as nuances do terreno ainda não foi testada no MDS nacional. Essas tecnologias têm um grande potencial para obtenção de MDE de elevada resolução e exatidão, visando ao levantamento preditivo de solos em pequenas áreas e/ou grandes escalas.

286 **Tabela 2** – Aspectos metodológicos do mapeamento digital de classes de solo no Brasil

Estudos	Total de amostras	Amostras /km²	Total amostrado (%)	Fonte do MDE	Resolução (m)	Software
Giasson et al. (2011)	1.333	217	0,54	foto	5	arcv, weka
ten Caten et al. (2011a)	70.000	80	20,00	srtm	50	arcg
ten Caten et al. (2011b)	58.440	67	16,72	srtm	50	arcg
ten Caten et al. (2011c)	70.000	80	20,00	srtm	50	arcg
Chagas et al.(2010)	3.000	303	27,27	cart	30	arcg, erda, java
Coelho & Giasson (2010)	11.000	11	8,75	srtm	90	arcg, weka
Crivelenti et al. (2009)	794.273	1.029	90,00	cart	30	arcg, ilwi, idri, weka
Carvalho et al.(2009)	424.643	1.111	100,00	srtm	30	arcg
Figueiredo & Giasson (2008)	7.200	10	8,10	srtm	90	arcv
Giasson et al. (2006)	7.500	30	25,00	srtm	92	arcv

287 foto: pares de fotografias estéreo, cart: cartas topográficas, srtm: shuttle radar topographic mission, arcg: ArcGIS
288 9, arcv: ArcView 3.2, ilwi: Ilwis, idri: Idrisi Andes, erda: Erdas, java: Java Neural Network Simulator, weka:
289 Weka
290

291 Predomina entre os estudos de MDS realizados no país a aplicação de softwares
292 proprietários. Os programas de código fonte fechados podem limitar a comunicação dos
293 procedimentos executados para a obtenção de determinado resultado (Wood, 2009). Entre as
294 alegações estaria o fato de que esses programas não possibilitam conhecer completamente as

rotinas e fórmulas empregadas nos processamento dos dados. Isso impossibilitaria a padronização de procedimentos de geração da informação. Existe uma variedade de programas com possíveis aplicações ao MDS que possuem seu código fonte aberto e uma grande rede de colaboradores para o desenvolvimento, a saber: SAGA, GRASS, ILWIS e o pacote estatístico R.

Qualidade e disponibilidade da informação

Quanto à avaliação da qualidade dos modelos preditivos, a matriz de erros e o índice kappa têm sido os procedimentos mais usuais (Tabela 3). A matriz de erros é uma boa alternativa para determinar a natureza e a frequência dos erros envolvidos. Contudo, caso a matriz seja gerada de maneira inapropriada, sem ser representativa dos dados, então a análise será sem significado (Congalton, 1991). O autor ressalta que os seguintes fatores precisam ser considerados: amostragem no campo, modo de obtenção do mapa a ser testado, correlação espacial, tamanho das amostras e esquema de amostragem.

Entre os estudos, em apenas dois casos foi avaliada a acurácia do mapeamento com dados de campo (Chagas et al., 2010; ten Caten et al., 2011b). Predomina entre os estudos apenas a avaliação do potencial de reprodução do mapa original a partir dos modelos preditivos. Essa metodologia compara o mapa predito com um mapa produzido por metodologias clássicas de levantamento de solos. Entretanto, pode não ser uma abordagem eficaz para indicar a qualidade do mapa de solos gerado, principalmente, devido às diferenças entre as metodologias empregadas.

Tabela 3 – Características dos estudos revisados quanto a aspectos de qualidade e disponibilidade da informação gerada.

Estudos	Qualidade	Índice kappa*	NC	Legenda simplificada	Idioma
Giasson et al. (2011)	mat / kap / não	0,52	2°	realizada	Ing
ten Caten et al. (2011a)	mat / kap / não	0,63**	ass, 1°, 2°	não realizada	Port
ten Caten et al. (2011b)	mat / kap / sim	0,46	ass, 2°	não realizada	Port
ten Caten et al. (2011c)	mat / kap / não	0,37	ass, 2°	não realizada	Port
Chagas et al.(2010)	mat / kap / sim	0,73	ass, 4°	realizada	Port
Coelho & Giasson (2010)	mat / kap / não	0,38	2°, 3°, 4°	realizada	Port
Crivelenti et al. (2009)	mat / kap / não	0,43	ass, 3°, 4°, textura	realizada	Port
Carvalho et al.(2009)	visual / não	-	ass, 2°, 3°, 4°	não realizada	Port
Figueiredo & Giasson (2008)	mat / kap / não	0,38***	ass, 3°, relevo, textura	realizada	Port
Giasson et al. (2006)	mat / kap / não	0,36***	ass, 1°, 2°	realizada	Ing

NC: nível categórico do SiBCS, mat: matriz de erros, kap: índice kappa, visual: comparação visual entre mapas, / sim: verificado no campo, / não: não verificado no campo, ass: associações, *: reprodutibilidade do mapa, **: dado não publicado, ***: sem simplificação da legenda, Ing: Inglês, Port: Português

O valor médio do índice kappa entre os estudos realizados no país é de 0,47 (Tabela 3). Esse valor é similar aos valores reportados na literatura internacional, como em Hengl & Rossiter (2003) de 0,58 em locais montanhosos e de 0,39 para áreas planas, e por Scull et al.

(2005) que observaram valores de 0,44 e 0,52 em áreas montanhosas e planas, respectivamente. A abordagem adotada em ambos os estudos, pela estratificação da área a ser mapeada de acordo com características locais predominantes, tem sido apresentada como uma possibilidade de melhorar o desempenho dos modelos preditivos, uma vez que a identificação e separação das áreas de acordo com critérios pré-estabelecidos é facilitada por sistemas de informação geográfica. Assim, possibilita-se a geração de modelos mais regionalizados, com melhor desempenho em cada uma das regiões para os quais foram desenvolvidos, permitindo-se capturar as peculiaridades de formação dos solos. A abordagem estratificada foi testada por Giasson et al. (2006), com um valor de kappa de 0,31 para a abordagem estratificada e de kappa de 0,36 quando não estratificada. Esses autores relataram que a possível causa do pior desempenho da abordagem estratificada tenha sido o modelo digital de elevação de 90 m utilizado no estudo.

A predição das classes de solos foi realizada nos quatro níveis categóricos do Sistema Brasileiro de Classificação de Solos (SiBCS). A quase totalidade dos estudos também buscou mapear classes de solos na forma de associações. Alguns estudos buscaram avaliar o potencial das metodologias utilizadas em mapear níveis categóricos mais baixos, como o 3° e o 4° níveis do SiBCS. No entanto, nesses níveis os solos são classificados de acordo com características de mais difícil associação com os fatores de formação do solo, a exemplo de características físicas e químicas no 3° nível e da presença de variações e características extraordinárias para a classificação no 4° nível. Acertos na distinção e na espacialização dessas classes na paisagem podem estar mais sendo fruto do acaso do que do verdadeiro poder preditivo dos modelos.

Alguns estudos têm se utilizado da estratégia de simplificar a legenda em relação a um mapa de referência, visando melhorar capacidade preditiva dos modelos. Essa abordagem consiste em mapear classes na forma de associações de solos ou no mapeamento de classes

apenas em níveis mais elevados do SiBCS. Tal estratégia melhora os indicadores de acerto em relação a um mapa de referência, contudo, produz mapas generalizados com um grau cada vez maior de variações dentro da unidade de mapeamento, quanto maior a simplificação adotada.

Os resultados das pesquisas em MDS realizadas no país são, na sua maioria, publicados em artigos escritos em língua portuguesa (Tabela 3). Como o Brasil é um país com uma diversidade de biomas e com grandes lacunas na cartografia de seus solos, os resultados aqui alcançados, sem dúvida, são de interesse da comunidade internacional. Entre as experiências já relatadas do MDS, predominam as condições de solos temperados e de regiões áridas do globo (McBratney et al., 2003; Scull et al., 2003; Grunwald, 2009). Faz-se fundamental que os estudos dessa linha de pesquisa realizados em condições de clima tropical sejam divulgados em outros países. A internacionalização das pesquisas em MDS realizadas no Brasil passa pela publicação dos resultados no idioma inglês.

Expectativas para o MDS no Brasil

Os custos do MDS não foram mencionados em nenhum dos estudos revisados. Para Grunwald (2009), a pesquisa relacionada aos investimentos necessários para tornar o MDS efetivo na produção de conhecimento espacial acerca de classes e propriedades de solos necessita mais atenção dos pesquisadores. Custos envolvidos no acesso a levantamentos de solos já existentes e coletas de novas informações, assim como dados oriundos de técnicas de sensoriamento proximal ou remoto deverão ser considerados. Além disso, conhecimento de especialistas em solos, estatística, informática e geomática precisa ser avaliado no custo total de projetos em MDS. O tempo investido em um projeto de MDS também deverá ser relatado em futuros estudos, para que novos projetos possam ser adequadamente planejados.

Uma padronização da abordagem para a avaliação da qualidade dos mapas gerados será fundamental. Na medida em que dados existentes (mapas de solos) conjuntamente com dados de apoio de fácil aquisição (SRTM), que possuem erros inerentes, são cada vez mais

utilizados para o MDS, é preciso que uma metodologia padronizada seja proposta para avaliar e possibilitar comparações entre resultados (Krol, 2008). Espera-se dos sistemas modernos para o mapeamento de solos mais qualidade do que dos mapeamentos convencionais, pois no MDS são empregados dados mais bem definidos e documentados em termos de posicionamento, atualização e integração de dados (Finke, 2007). Para esse autor, é necessário estabelecer padrões de qualidade em termos da conformidade da informação com a realidade do campo (acurácia) e com a incerteza associada à metodologia automatizada que se utiliza para gerar a informação (precisão).

A avaliação da qualidade dos mapas de classes de solos a partir dos resultados da matriz de erros e índice kappa em relação ao mapa original de solos pode estar subestimando a verdadeira qualidade dos mapas gerados pelo MDS. Como os mapas convencionais são a representação cartográfica do conhecimento pedológico empregado na sua produção, e o conhecimento tácito do pedólogo está em constante aprimoramento, haverá sempre distorções entre o mapa gerado pelo pedólogo e a verdadeira classe de solo presente na paisagem. De tal forma que a comparação do mapa gerado pela metodologia automatizada com aquele obtido pela metodologia convencional apenas indica a capacidade de reprodução desse último pelo modelo preditivo. A checagem do mapa predito com informações de campo deve ser o verdadeiro indicador de qualidade a ser empregado.

Como uma das prerrogativas para a execução de um projeto de MDS é a sua implementação em um sistema de informação geográfica, a disponibilização da informação gerada em formatos vetorial ou matricial é facilitada. Contudo, não é o que tem sido relatado nos estudos realizados no país. Em nenhum dos artigos revisados, os autores relatam a disponibilização dos mapas gerados. Além de a forma digital possibilitar a rápida atualização dos mapas e sua disponibilização para acesso pela internet, essa abordagem permite que metodologias inovadoras de acesso, como visualização 3D (Grunwald et al., 2007) e

tecnologias móveis (iPhone) (Beaudette & O'Geen, 2010), sejam utilizadas. Essas estratégias permitirão maior divulgação das informações geradas. Um passo que precisa ser dado pelos pesquisadores brasileiros em MDS.

CONSIDERAÇÕES FINAIS

Os pedólogos têm se questionado sobre as motivações para a redução da valorização de sua atividade profissional. Entre os culpados, são elencadas a valorização da pesquisa aplicada em detrimento da básica, os critérios taxonômicos utilizados pela pedologia, os planejadores sem interesse pela atividade do pedólogo, levantamentos pedológicos insuficientes para aplicação imediata em problemas da sociedade (Ker & Novais, 2003). Em contraste com a argumentada recessão verificada em relação aos levantamentos de solos, verifica-se uma tendência de renovação, principalmente, quanto às técnicas de coleta e processamento de dados, formação e desenvolvimento de modelos e ao MDS (Ramos, 2003).

Face ao grande desafio de produzir informação sobre a distribuição espacial de classes de solos para todo o território brasileiro, o mapeamento digital de solos será um grande aliado. A abordagem tem se consolidado com um número cada vez maior de artigos publicados em revistas científicas especializadas, bem como com a participação de pesquisadores em publicações internacionais.

Há a necessidade de que sejam somados esforços para o treinamento de novos pedólogos aptos a implementar projetos de MDS, bem como em realizar levantamentos de solos por meio das metodologias convencionais. Ambas as abordagens são complementares e necessitam serem desenvolvidas simultaneamente. O levantamento de solos convencional produz as informações que são utilizadas para treinar os modelos e predizer classes e propriedades de solos em áreas não mapeadas. A abordagem automatizada permite que extensas regiões sejam previamente mapeadas, otimizando recursos humanos e financeiros no mapeamento convencional. Além disso, possibilita-se a geração de informação em formatos

vetorial e matricial prontamente disponível para projetos ambientais e de adequabilidade de uso.

O MDS brasileiro precisa experimentar-se mais na geração de informações sobre propriedades do solo. O mapeamento de classes de solos atende uma gama muito grande de aplicações. Contudo, o mapeamento de propriedades pode ser executado em uma abordagem *demand driven*, focando em clientes e necessidades específicas. Essa abordagem irá mobilizar os recursos estritamente necessários para o conhecimento de uma determinada propriedade do solo, por exemplo, para a predição da matéria orgânica do solo, dispensando a descrição completa de perfis. Também pode ocorrer por meio de verificação a campo, podendo-se gerar mapas de precisão e acurácia associados, indicando o grau de confiabilidade dos dados a serem disponibilizados.

REFERÊNCIAS

BALLABIO, C. Spatial prediction of soil properties in temperate mountain regions using support vector regression, **Geoderma**, v.151, n.3-4, p.338-350, 2009. Disponível em: http://www.sciencedirect.com/science/article/pii/S0016706109001499. Acesso em: 12 out. 2011. doi: 10.1016/j.geoderma.2009.04.022.

BEAUDETTE D.E.; O'GEEN A. T. An iPhone Application for On-Demand Access to Digital Soil Survey Information. **Soil Science Society of America Journal.** v.74, n.5, p.1682-1684, 2010. Disponível em: http://casoilresource.lawr.ucdavis.edu/dylan/publications/sssaj-iphone_app.pdf Acesso em: 12 out. 2011. doi: 10.2136/sssaj2010.0144N.

BURROUGH, P.A.; BOUMA, J.; YATES, S.R. The state of the art in pedometrics. **Geoderma**, v.62, p.311-326, 1994. Disponível em: http://www.sciencedirect.com/science/article/pii/0016706194900434 Acesso em: 12 out. 2011. doi: 10.1016/0016-7061(94)90043-4.

455 CARRE, F. et al. Digital soil assessments: Beyond DSM, **Geoderma**, v.142, n.1-2, p.69-79,
456 2007. Disponível em: http://www.sciencedirect.com/science/article/pii/S0016706107002261
457 Acesso em: 12 out. 2011. doi: 10.1016/j.geoderma.2007.08.015.
458 CARVALHO, C.C.N. et al. Mapa digital de solos: Uma proposta metodológica usando
459 inferência fuzzy. **Revista Brasileira de Engenharia Agrícola e Ambiental**, v.13, n.1, p.46-
460 55, 2009. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1415-
461 436620090 00100007&lng=en&nrm=iso. Acesso em: 25 mai. 2011. doi: 10.1590/S1415-
462 43662009000100007.
463 CHAGAS, C. da S. et al. Atributos topográficos e dados do Landsat7 no mapeamento digital
464 de solos com uso de redes neurais. **Pesquisa Agropecuária Brasileira**, v.45, n.5, p.497-507,
465 2010. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1415-
466 43662009000100007&lng=en&nrm=iso. Acesso em: 25 mai. 2011. 10.1590/S1415-
467 43662009000100007.
468 COELHO, F.F.; GIASSON, E. Comparação de métodos para mapeamento digital de solos
469 com utilização de sistema de informação geográfica. **Ciência Rural**, v.40, n.10, p.2099-
470 2106, 2010. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-
471 84782010001000008&lng=en&nrm=iso. Acesso em: 20 fev. 2011. doi: 10.1590/S0103-
472 84782010005000156.
473 CONGALTON, R.G. A Review of Assessing the Accuracy of Classifications of Remotely
474 Sensed Data. **Remote Sensing of Environment**, v.37, p.35-46, 1991. Disponível em:
475 http://www.sciencedirect.com/science/article/pii/003442579190048B Acesso em: 12 out.
476 2011. doi: 10.1016/0034-4257(91)90048-B.
477 CPRM – Serviço geológico do Brasil. Disponível em: http://www.cprm.gov.br/ Acessado em
478 10 ago. 2011.

479 CRIVELENTI et al. Mineração de dados para a inferência de relações solo-paisagem em
480 mapeamentos digitais de solo. **Revista Agropecuária Brasileira**, v.44, n.12, p.1707-1715,
481 2009. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-
482 204X2009001200021&lng=en&nrm=iso. Acesso em: 25 mai. 2011. doi: 10.1590/S0100-
483 204X2009001200021.
484 FIGUEIREDO, S.R. et al. Uso de regressões logísticas múltiplas para mapeamento digital de
485 solos no planalto médio do RS. **Revista Brasileira de Ciência do Solo**, v.32, p.2779-2785,
486 2008. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-
487 06832008000700023&lng=en&nrm=iso. Acesso em: 20 fev. 2011. doi: 10.1590/S0100-
488 06832008000700023.
489 FINKE, P.A. Quality assessment of digital soil maps: producers and users perspectives. In:
490 LAGACHERIE, P.; MCBRATNEY, A.B.; VOLTZ, M., Digital Soil Mapping: An
491 Introductory Perspective. Developments in Soil Science, v.31, cap.39, Elsevier, Amsterdam,
492 2007, p.523-541.
493 HARTEMINK, A.E. et al. Digital Soil Mapping with Limited Data. Ed. Springer, 2008. 445p.
494 HARTEMINK, A.E.; MCBRATNEY, A.B., A soil science renaissance. **Geoderma**, v.148,
495 n.2, p.123-129, 2008. Disponível em:
496 http://www.sciencedirect.com/science/article/pii/S0016706108002802 Acesso em: 12 out.
497 2011. doi: 10.1016/j.geoderma.2008.10.006.
498 HENGL, T.; ROSSITER, D.G. Supervised Landform classification to enhance and replace
499 photo-interpretation in semi-detailed soil survey. **Soil Science Society of America Journal**.
500 v.67, p.1810-1822, 2003. Disponível em:
501 https://www.soils.org/publications/sssaj/abstracts/67/6/1810. Acesso em: 14 nov. 2011. doi:
502 10.2136/sssaj2003.1810.

503 GIASSON, E. et al. Decision trees for digital soil mapping on subtropical basaltic steeplands.
504 **Scientia Agrícola,** v.68, p.167-174, 2011. Disponível em:
505 http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-90162011000200006&lng=en
506 &nrm=iso. Acesso em: 25 mai. 2011. doi: 10.1590/S0103-90162011000200006.
507 GIASSON, E. et al. Digital soil mapping using multiple logistic regression on terrain
508 parameters in southern Brazil. **Scientia Agricola**, v.63, p.262-268, 2006. Disponível em:
509 http://www.scielo.br/scielo.php?script=sciarttext&pid=S0103-90162006000300008&lng=en
510 & nrm=iso. Acesso em: 20 fev. 2011. doi: 10.1590/S0103-90162006000300008.
511 GRUNWALD S. Multi-criteria characterization of recent digital soil mapping and modeling
512 approaches. **Geoderma**, v.152, p.195-207. 2009. Disponível em:
513 http://www.sciencedirect.com/science/article/pii/S0016706109001827 Acesso em: 12 out.
514 2011. doi: 10.1016/j.geoderma.2009.06.003.
515 GRUNWALD, S. et al. Are Current Scientific Visualization and Virtual Reality Techniques
516 Capable to Represent Real Soil-Landscapes? In: LAGACHERIE, P.; MCBRATNEY, A.B.;
517 VOLTZ, M., Digital Soil Mapping: An Introductory Perspective. Developments in Soil
518 Science, v.31, cap.42, Elsevier, Amsterdam, 2007, p.571-580.
519 LAGACHERIE, P. & MCBRATNEY, A. B. Spatial soil information systems and spatial soil
520 inference systems: perspectives for digital soil mapping. In: LAGACHERIE, P.;
521 MCBRATNEY, A. & VOLTZ, M. ed. **Digital soil mapping:** an introductory perspective.
522 Amsterdam: Elsevier, 2007, p. 3-22.
523 KEMPEN, B.; et al. Updating the 1:50,000 Dutch soil map using legacy soil data: a
524 multinomial logistic regression approach. **Geoderma**, v.151, p.311‑326, 2009. Disponível
525 em: http://www.sciencedirect.com/science/article/pii/S0016706109001475 Acesso em: 20
526 fev. 2011. doi: 10.1016/j.geoderma.2009.04.023.

527 KHEIR, R.B. et al. Predictive mapping of soil organic carbon in wet cultivated lands using
528 classification-tree based models: The case study of Denmark. **Journal of Environmental**
529 **Management**. v.91, n.5, p. 1150-1160. 2010a. Disponível em:
530 http://www.sciencedirect.com/science/article/pii/S0301479710000022 Acesso em: 12 out.
531 2011. doi: 10.1016/j.jenvman.2010.01.001.
532 KHEIR, R.B. et al. Spatial soil zinc content distribution from terrain parameters: A GIS-based
533 decision-tree model in Lebanon. **Environmental Pollution**, v.158, n.2, p.520-528, 2010b.
534 Disponível em: http://www.sciencedirect.com/science/article/pii/S0269749109004163
535 Acesso em: 12 out. 2011. doi: 10.1016/j.envpol.2009.08.009.
536 KER, J.C. ; NOVAIS, R.F. Fundamentos da pedologia e relação com a fertilidade do solo. In:
537 XXIX Congresso Brasileiro de Ciência do Solo, 2003, Ribeirão Preto, 2003.
538 KROL, B.G.C.M. Towards a Data Quality Management Framework for Digital Soil Mapping
539 with Limited Data. In: Digital Soil Mapping with Limited Data. HARTEMINK, A.E.;
540 MCBRATNEY, A.B.; MENDONÇA-SANTOS, M. de L. (Eds.), cap.11, Springer, 2008,
541 p.137-149.
542 MCBRATNEY, A. B.; MENDONCA SANTOS, M. L. & MINASNY, B. On digital soil
543 mapping. **Geoderma**, v.17, p.3-52, 2003. Disponível em:
544 http://www.sciencedirect.com/science/article/pii/S0016706103002234. Acesso em: 12 0ut.
545 2011. doi: 10.1016/S0016-7061(03)00223-4.
546 MENDONÇA-SANTOS, M.L. & SANTOS, H.G. dos The state of the art of brazilian soil
547 mapping and prospects for digital soil mapping. In: Digital soil mapping: an introductory
548 perspective. LAGACHERIE P., MCBRATNEY A.B. & VOLTZ, M. (ed.), cap.3, p.39-54,
549 2007.
550 MINASNY, B. et al. Digital Soil Mapping Technologies for Countries with Sparse Data
551 Infrastructures. In: Digital Soil Mapping With Limited Data. HARTEMINK, A.E.;

MCBRATNEY, A.B.; MENDONÇA-SANTOS, M. de L. (Eds.), cap.2, Springer. 2008, p.15-30.

QI, F.; ZHU, A. X. Knowledge discovery from soil maps using inductive learning. **International Journal of Geographical Information Science**. v.17, n.8, p.771-795, 2003. Disponível em: http://solimserver.geography.wisc.edu/pdfs/QiFeng_IJGIS2003.pdf Acesso em: 12 out. 2011. doi: 10.1080/1365881031000159604 9.

RAMOS D.P. Desafios da Pedologia Brasileira frente ao novo milênio - (CNPS/EMBRAPA). In: Simpósio do XXIX Congresso Brasileiro de Ciência do Solo, Ribeirão Preto, SP, 2003.

SACHS, J. et al. Monitoring the world's agriculture. **Nature**, v.466, p.558–560, 2010. Disponível em: http://www.css.cornell.edu/faculty/lehmann/publ/Nature%20466,%20558-560,%202010%20Sachs.pdf Acesso em: 12 out. 2011. doi:10.1038/466558a.

SANCHEZ, P.A. et al. Digital soil map of the world. **Science**, v.325, p.680-681, 2009. Disponível em: http://www.sciencemag.org/content/325/5941/680.full.pdf Acesso em: 12 out. 2011. doi: 10.1126/science.1175084.

SARMENTO, E. C. et al. Sistema de informação geográfica como apoio ao levantamento detalhado de solos do Vale dos Vinhedos. **Revista Brasileira de Ciência do Solo**, v.32, p.2795-2803. 2008. Disponível em <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-06832008000700025&lng=en&nrm=iso>. Acesso em: 23 ago. 2011. doi: http://dx.org/ 10.1590/S0100-06832008000700025.

SCHMIDT, K. et al. Instance selection and classification tree analysis for large spatial datasets in digital soil mapping. **Geoderma**, v.146, p.138-146, 2008. Disponível em: http://www.sciencedirect.com/science/article/pii/S0016706108001304 Acesso em: 12 out. 2011. doi: 10.1016/j.geoderma.2008.05.010.

575 SCULL, P. et al. Predictive soil mapping: a review. **Progress in Physical Geography**, v.27,
576 p.171-197, 2003. Disponível em: http://ppg.sagepub.com/content/27/2/171. Acesso em: 14
577 nov. 2011. doi: 10.1191/0309133303pp366ra.
578 SCULL, P.; et al. The application of classification tree analysis to soil type prediction in a
579 desert landscape. **Ecological Modelling**, v.181, p.1-15, 2005. Disponível em:
580 http://www.sciencedirect.com/science/article/pii/S0304380004003540. Acesso em: 14 nov.
581 2011. doi: 10.1016/j.ecolmodel.2004.06.036.
582 TEN CATEN, A. et al. Regressões logísticas múltiplas: fatores que influenciam sua aplicação
583 na predição de classes de solos. **Revista Brasileira de Ciência do Solo**, v.35, n.1, p.53-62,
584 2011a. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-
585 068320_11000100005&lng=en&nrm=iso. Acesso em: 16 mai. 2011. doi: 10.1590/S0100-
586 06832011000100005.
587 TEN CATEN, A.; et al. Extrapolação das relações solo-paisagem a partir de uma área de
588 referência. **Ciência Rural**, v.41, n.5, p. 812-816, 2011b. Disponível em:
589 http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-84782011000500012&lng=
590 en&nrm=iso. Acesso em: 12 out. 2011. doi: 10.1590/S0103-84782011000500012
591 TEN CATEN et al. Componentes principais como preditores no mapeamento digital de
592 classes de solos. **Ciência Rural,** v.41, n.7, p. 1170-1176, 2011c. Disponível em:
593 http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-84782011000700011&lng=
594 en&nrm=iso. Acesso em: 12 out. 2011. doi: 10.1590/S0103-84782011000700011.
595 VALERIANO, M. M. ; ROSSETTI, D. F. Topodata: seleção de coeficientes geoestatísticos
596 para refinamento unificado de dados SRTM. São José dos Campos, SP: NPE: Coordenação de
597 Ensino, Documentação e Programas Especiais (INPE-16701-RPQ/853) (Boletim). 2010. 74p.
598 WEBSTER, R.; BURROUGH, P.A. Computer-based soil mapping of small areas from
599 sample data. 1. Multivariate classification and ordination. **Journal of Soil Science**, v.23,

600 p.210-221, 1972. Disponível em: http://dx.doi.org/10.1111/j.1365-2389.1972.tb01654.x
601 Acesso em: 12 out. 2011. doi.: 10.1111/j.1365-2389.1972.tb01654.x.
602 WITTEN, I.H. & FRANK, E. Data Mining: Practical Machine Learning Tools and
603 Techniques, 2nd ed., Morgan Kaufmann, 2005, 560p.
604 WOOD, J. Overview of Software Packages Used in Geomorphometry. Cap.10. In.:
605 Geomorphometry: Concepts, Software, Applications. HENGL, T. & REUTER, H.I. (ed).
606 Elsevier, 2009, 76.

ARTIGO 2 - Resolução espacial de um modelo digital de elevação definida pela função wavelet.[(1)]

Resumo – O objetivo do presente estudo foi empregar a transformada wavelet para avaliar a variabilidade de quatro atributos de terreno, possibilitando definir a resolução espacial mais apropriada para representar a variabilidade desses. Os dados utilizados no estudo têm sua origem em três transetos de 27 km posicionados em áreas do Planalto, Rebordo do Planalto e Depressão Central, na região central do Estado do Rio Grande do Sul. Foram utilizados os atributos elevação, declividade, curvatura em perfil e índice de umidade topográfica. Tais atributos foram derivados de um modelo digital de elevação Topodata com resolução de 30 m. A avaliação da resolução em que ocorria a máxima variabilidade foi realizada pela aplicação da wavelet mãe denominada Morlet. Os resultados foram avaliados a partir do isograma e do escalograma dos coeficientes wavelets e indicam que sensores remotos com resolução espacial próxima a 60 e 100 m podem ser utilizados em pesquisas que considerem os atributos de terreno declividade, curvatura em perfil e índice de umidade topográfica, ou ainda, fenômenos ambientais correlacionados a esses. No entanto, não foi possível estabelecer um valor conclusivo para a resolução espacial mais adequada para o atributo elevação.

Termos para indexação: mapeamento digital de solos, wavelet mãe de Morlet, escala, pedometria.

Spatial resolution of a digital elevation model defined by the wavelet function

Abstract – The aim of this study was to use wavelet transform to assess the variability of four terrain attributes, allowing the appropriate definition of spatial resolution to represent the variability of each attribute. The data used in the study had their origin in three 27 km transects positioned in areas of the Planalto, Rebordo do Planalto and Depressão Central in the central region of the Rio Grande do Sul State. The attributes elevation, slope, profile

curvature and topographic wetness index were derived from a Topodata digital elevation model with resolution of 30 m. The assessment of the resolution with the highest spatial variability was evaluated trough the application of the Morlet mother wavelet. The results were evaluated from wavelet coefficients plotted in isograms and scalograms. The study results indicate that remote sensors with spatial resolution between 60 and 100 m should be used in researches considering the terrain attributes slope, profile curvature and topographic wetness index, or even environmental phenomena related to them. The study was unable to establish a conclusive value for a more appropriate spatial resolution to the attribute elevation.
Index terms: digital soil mapping, Morlet mother wavelet, scale, pedometrics.

INTRODUÇÃO

Variáveis ambientais normalmente apresentam uma substancial variabilidade em uma ampla gama de escalas espaciais. As propriedades do solo estão entre as variáveis que, em geral, exibem um complexo padrão espacial fruto das interações entre processos físicos, químicos e biológicos, os quais ocorrem em uma variedade de escalas espaciais. A elucidação das escalas em que esses processos se desenvolvem é essencial para o entendimento e a predição de processos hidrológicos, biológicos e químicos, que atuam diretamente na gênese do solo (Si, 2008).

A resolução espacial está intimamente relacionada à escala (Bian, 1997; Hengl, 2006). Entre as dúvidas dos pesquisadores está a que diz respeito à resolução espacial em que os modelos baseados em parâmetros de terreno irão desempenhar seu máximo. Com a diminuição da resolução espacial, os parâmetros primários de terreno derivados do modelo digital de elevação (MDE) perdem detalhamento, e muitas nuanças da paisagem também são perdidas (Wu et al., 2008a).

Segundo Wu et al. (2008b), diferentes resoluções espaciais do MDE produzem distintas informações sobre os atributos de terreno. À medida que o MDE é agregado em resoluções

cada vez mais grosseiras, a declividade diminui, e a área de contribuição aumenta. De acordo com os autores, existe um valor ótimo para a resolução espacial, a partir do qual o modelo hidrológico irá gerar os melhores resultados. Embora a resolução espacial dependa da complexidade da paisagem e dos atributos de terreno utilizados na modelagem, Hengl (2006) argumenta que muitos trabalhos, os quais utilizam dados em formato matricial (raster), optam por uma determinada resolução espacial sem qualquer justificativa científica em favor da opção adotada.

A estratégia automatizada de mapeamento do solo, denominada Mapeamento Digital de Solos (MDS), emprega preditores derivados de MDE para a predição espacial de classes e propriedades do solo (Sanchez et al., 2009). Um conjunto de preditores de larga aplicação no MDS são os atributos de terreno (Scull et al., 2003). Esses podem ser derivados diretamente de MDE, como é o caso da declividade, direção do fluxo e curvaturas. Outros atributos são de definição um pouco mais complexa e constituem-se em atributos mais elaborados resultados da análise geomorfométricas da paisagem. Esse é o caso do índice de posição topográfica (IPT), que é utilizado para classificar a paisagem em classes morfológicas (Tagil & Jenness, 2008).

A metodologia utilizada influencia sobremaneira na qualidade dos valores gerados quando da derivação do IPT (Judex et al., 2006). O IPT representa a diferença entre a elevação de um pixel e a elevação média dos pixels vizinhos, possibilitando uma classificação simples e rápida da paisagem em classes morfológicas. Como esse índice está relacionado à extensão e à forma com que o algoritmo analisa pixels vizinhos, os parâmetros adotados durante a geração do índice devem ser produto de uma criteriosa análise prévia das características geomorfológicas da região para a qual os dados estão sendo gerados. Com isso, será possível avaliar a extensão e os padrões da variabilidade espacial, os quais têm reflexo direto na resolução espacial de trabalho (Tagil & Jenness, 2008).

Entre as alternativas de estudo das resoluções espaciais mais apropriadas para os diversos fenômenos ambientais, tem-se a aplicação da análise wavelet (Cho & Chon, 2006; Dong et al., 2008). A aplicação da análise wavelet consiste em mimetizar os dados originais a partir de um sinal artificial produzido por uma pequena onda, a wavelet mãe. A wavelet é dilatada e transladada a fim de buscar a melhor convolução entre os dados originais e a wavelet mãe. Desse procedimento são gerados os coeficientes wavelet, os quais representam o grau de correlação entre o sinal original e o sinal gerado. Esses coeficientes podem, então, ser visualizados na forma do isograma local e escalograma global, que irão auxiliar na avaliação das estruturas predominantes nos dados em questão (Torrence & Compo, 1998).

A wavelet é uma função matemática representada na forma de uma pequena onda. É considerada pequena, porque contrasta com as funções seno e cosseno típicas da análise de Fourier que se estendem para o infinito. A análise wavelet é executada por meio da transformada wavelet, a qual decompõe a variância de um processo físico em uma série de coeficientes que representam a distribuição da variância em diferentes resoluções espaciais e posições da paisagem (Yates et al., 2006). A análise wavelet constitui-se em uma boa alternativa para a análise de dados assimétricos e não estacionários comuns em dados ambientais (Lark & Webster, 1999).

Este estudo teve como objetivo empregar a transformada wavelet para avaliar a variabilidade dos atributos de terreno elevação, declividade, curvatura em perfil e índice de umidade topográfica em três regiões fisiográficas distintas do Estado do Rio Grande do Sul. O estudo buscou definir a resolução espacial mais apropriada para representar a variabilidade de cada um dos quatro atributos nessas regiões, além de demonstrar a aplicação da análise wavelet, que se constitui como uma ferramenta com enorme potencial de aplicação para estudos espaço-temporais.

MATERIAL E MÉTODOS

Área de estudo

Os dados utilizados no estudo têm sua origem em três transetos posicionados em distintas regiões fisiográficas típicas da região central do Estado do Rio Grande do Sul (Figura 1). Essas três regiões são o Planalto da Serra Geral, o Rebordo do Planalto e a Depressão Central. A complexidade geomorfológica dessa região do Estado implica em um desafio para o mapeamento preditivo de classes e propriedades do solo devido à diversidade de nuanças do relevo que, por sua vez, influencia sobremaneira a formação dos solos da região. Cada transeto tem um comprimento de aproximadamente 27 quilômetros, o que possibilitou, na atual resolução espacial dos atributos de terreno, um total de 900 pixels amostrados por transeto, bem superior às 128 amostras preconizadas como valor mínimo para o emprego da análise wavelet (Si, 2008).

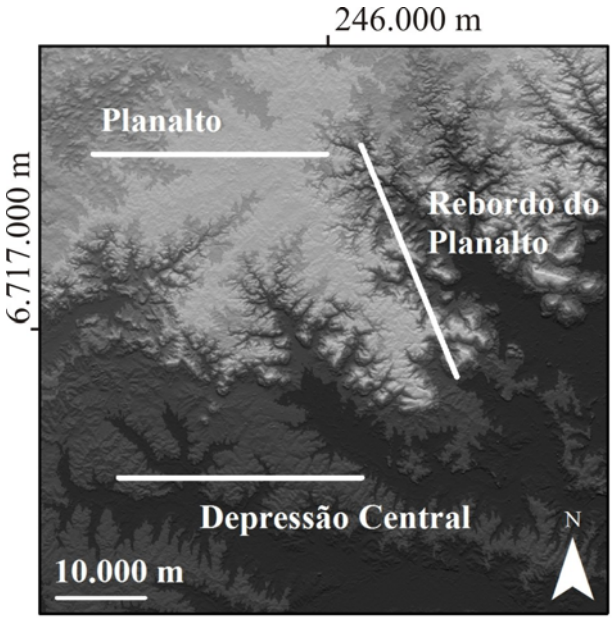


Figura 1. Localização dos transetos em três regiões fisiográficas predominantes da região central do Estado do Rio Grande do Sul. Níveis de cinza claros representam cotas acima de 400 metros, níveis escuros identificam locais com cota em torno de 60 m. Coordenadas UTM referenciadas ao datum SIRGAS 2000/Fuso22.

Atributos de terreno

Optou-se por utilizar atributos de terreno com significado físico e com relação à formação dos solos, visando a sua aplicação posterior em estudos relacionados ao MDS. A elevação, ou altitude acima de uma superfície de referência, tem importante papel na definição do clima local. A declividade, como medida da taxa de mudança da elevação na direção do declive, afeta a velocidade de fluxos superficiais e subsuperficiais. A curvatura em perfil, que representa a taxa de mudança da declividade na direção de uma linha de fluxo, é fundamental no comportamento da água sobre a paisagem. O índice de umidade topográfica, que é uma função do logaritmo natural da razão entre área de contribuição e a declividade, indica o controle da topografia sobre a umidade do solo (Wilson & Gallant, 2000).

Os quatro atributos foram derivados do MDE e extraídos do Topodata (Valeriano & Rossetti, 2010), que, por sua vez, consiste em uma interpolação dos dados SRTM (*Shuttle Radar Topography Mission*). Esse MDE apresenta resolução espacial relativamente elevada (30 m) e é obtido por uma única técnica, além de estar disponível de maneira contínua para extensas regiões. Os arquivos raster de atributos de terreno foram gerados a partir do programa TAPES-G de acordo com procedimento detalhado por Wilson & Gallant (2000).

Cada um dos atributos de terreno foi amostrado a partir do respectivo transeto pela função "criar gráfico de perfil" do programa ArcMap 9.3. Essa função permite a exportação direta de todos os valores dos pixels encobertos pela linha do transeto imediatamente para o formato ASCII.

Análise wavelet

A análise foi executada por meio da transformada wavelet, como segue:

$$W(b,x_n)=\frac{1}{\sqrt{b}}\sum_{j=1}^{n} f(x_j)g\left(\frac{x_j-x_n}{b}\right)$$

onde: $W(b,x_n)$ refere-se aos coeficientes wavelet, positivos ou negativos de acordo com a correlação entre a wavelet e o sinal original; j representa a j-éssima observação; n é o total de

observações; b representa a resolução que está sendo testada centrada em x_n; x_j é a posição onde a análise está sendo executada no transeto. A função $f(x_j)$ representa os dados originais, e a função $g(x_j)$ é a wavelet mãe. Neste estudo, foi utilizada a wavelet mãe denominada Morlet, como segue:

$$g(x_j) = \pi^{-\left(\frac{1}{4}\right)} e^{i6\eta} e^{-\left(\frac{1}{2}\right)\eta^2}$$

em que

$\eta = \frac{b}{x_j}$ (Torrence & Compo, 1998).

Durante a transformada wavelet, valores elevados de variância indicam que houve um bom ajuste entre a wavelet e os dados originais. Esse processo é repetido para uma faixa de resoluções ao longo de todo o transeto, dando origem ao isograma local (*wavelet power spectrum*), em que os coeficientes wavelet são plotados ao quadrado, resultando apenas valores positivos (Torrence & Compo, 1998).

Os resultados da análise wavelet são altamente correlacionados com a escolha da wavelet mãe mais apropriada. Tal ponto merece uma especial atenção (Labat, 2005). A wavelet do tipo Morlet foi utilizada neste estudo devido à sua superior capacidade de detectar e localizar padrões nos dados (Mi et al., 2005).

A variância total v(b) foi calculada por:

$$v(b) = \frac{1}{n}\sum_{i=1}^{n} W^2(b, x_n)$$

A variância total é importante na medida em que seu valor será máximo quando ocorrer a melhor convolução entre as funções $f(x_j)$ e $g(x_j)$, ou seja, a variância será maximizada quando a resolução for igual ao tamanho médio das estruturas, como fragmentos e interrupções presentes nos dados (Si, 2008). O resultado da variância total será representado no escalograma global (*global wavelet power*). O espectro de ruído vermelho (Torrence &

Compo, 1998) ao nível de significância de 5% foi utilizado como hipótese nula para o teste de significância das informações presentes no isograma e no escalograma. Dessa forma, regiões do isograma local circundadas por isolinhas denotam regiões de elevada correlação entre o sinal original e a wavelet mãe, o que indicará padrões relevantes nos dados. No caso do escalograma, locais onde a variância total é superior à linha de significância, as resoluções espaciais são relevantes para o presente estudo.

Sobre os isogramas dos coeficientes wavelets, foi representado o cone de influência. Os cones de influência delimitam a presença de efeitos de bordas nos dados próximos às extremidades de cada transeto. Informações nessa região dos isogramas não devem ser consideradas para efeito de afirmativas sobre o padrão de resoluções presentes nos dados originais (Torrence & Compo, 1998). A análise wavelet dos dados foi executada de acordo com o algoritmo descrito e disponibilizado por Torrence & Compo (1998).

Gráficos de variância (isogramas e escalogramas) foram utilizados para avaliar os padrões de resolução para cada atributo em cada transeto. Partindo-se do pressuposto de que as áreas do Planalto e da Depressão Central (Figura 1) constituem-se em áreas morfologicamente mais similares, os gráficos de variância foram agrupados em quatro figuras, considerando-se um atributo em duas localidades distintas, a saber: elevação no Planalto e no Rebordo, curvatura em perfil no Planalto e no Rebordo, declividade na Depressão e no Rebordo e índice de umidade topográfica na Depressão e no Rebordo. Dessa forma, são executadas comparações em um mesmo atributo entre duas áreas de maior contrastante, além de possibilitar a análise dos quatro atributos na área do rebordo, onde há a maior diversidade de formas na paisagem.

Para definir a resolução espacial (tamanho de pixel) que melhor representa a variabilidade espacial existente nos quatro atributos, foi adotado o critério de He et al. (2007). Para esses autores, a resolução espacial do pixel deve ser de um quarto (1/4) do valor identificado pela

189 análise wavelet. Segundo os mesmos autores, com esse critério, há uma maior segurança para
190 evitar que padrões importantes associados com os atributos não sejam perdidos.
191
192 **RESULTADOS E DISCUSSÃO**
193 **Atributos de terreno**
194 A análise preliminar dos dados advindos da amostragem nos três transetos permite
195 visualizar características típicas das regiões fisiográficas onde os dados foram coletados
196 (Tabela 1). O menor (115,329 m) e o maior (491,970 m) valor para o atributo elevação, no
197 transeto localizado na área do Rebordo do Planalto, são reflexo da conformação do relevo
198 nessa área. Localizada na transição entre o Planalto Meridional Brasileiro e a Depressão
199 Periférica, essa região tem como características a presença de paisagem fortemente dissecada,
200 formando vales em forma de "V". Essas formas são resultado dos processos erosivos que ao
201 longo de milhares de anos vêm esculpindo a região, gerando esse padrão de grande
202 variabilidade na geomorfologia (Uberti & Klamt, 1984). O desvio padrão desse atributo para
203 o Planalto (σ = 32,329 m) e para a Depressão Central (σ = 20,352 m) indica que essas áreas
204 são mais planas do que se comparadas a área do Rebordo do Planalto.
205 Quanto à declividade, o valor máximo (41,963 °), o valor médio (13,720 °) e o desvio
206 padrão (9,830 °) desse atributo em áreas do Rebordo indicam relevo acidentado. Comparando
207 os valores da declividade no Rebordo com os obtidos no Planalto e Depressão, verifica-se que
208 nessas áreas o relevo é mais suave. Segundo Sartori (2009), na Depressão Central
209 predominam as colinas e planícies aluviais formadas a partir de rochas sedimentares. Nas
210 áreas do Planalto, o relevo varia de ondulado a suave ondulado formado sobre uma sucessão
211 de derrames vulcânicos sobrepostos.
212 A curvatura em perfil indica a taxa de mudança do relevo ao longo do declive, valores
213 positivos indicam superfícies convexas com possível perda de material, que, por sua vez,
214 acumula-se nos locais côncavos com curvatura negativa (Wilson & Gallant, 2000). As áreas

de Rebordo possuem as maiores curvaturas entre os três transetos (Tabela 1), o que era esperado devido à complexidade do terreno. A suavidade do relevo na Depressão e Planalto é retratada pelo pequeno desvio padrão do atributo curvatura em perfil.

Tabela 1. Estatística descritiva dos quatro parâmetros de terreno nos três transetos utilizados no estudo.

Parâmetro de terreno	Transeto	Mínimo	Máximo	Média	Mediana	Desvio Padrão	Variância
Elevação (m)	Rebordo	115,329	491,970	272,617	240,220	106,159	11269,818
	Planalto	383,259	513,487	462,471	475,019	32,329	1045,153
	Depressão	57,715	140,803	97,904	100,173	20,352	414,187
Curvatura em perfil (m^{-1})	Rebordo	-1,241	0,996	0,015	0,028	0,221	0,049
	Planalto	-0,296	0,253	0,005	0,007	0,063	0,004
	Depressão	-0,182	0,450	0,004	-0,001	0,062	0,004
Declividade (°)	Rebordo	0,240	41,963	13,720	10,854	9,830	96,639
	Planalto	0,264	17,088	3,382	2,946	2,047	4,188
	Depressão	0,342	9,547	3,430	3,203	1,946	3,786
Índice de umidade topográfica	Rebordo	4,657	16,306	7,458	6,869	1,985	3,942
	Planalto	5,734	18,725	8,553	8,193	1,524	2,322
	Depressão	6,119	19,531	8,467	7,987	1,569	2,462

O índice de umidade topográfica, por se tratar de um atributo que considera a área de contribuição da bacia e a declividade do terreno, possibilita maior semelhança entre os valores do Planalto, Rebordo e Depressão (Tabela 1). Contudo, as maiores amplitudes entre áreas úmidas (valores elevados) e áreas drenadas (valores baixos) foram verificadas na Depressão. Para essa região, drenam todos os córregos que se originam nas encostas do Rebordo do Planalto (Sartori, 2009), possibilitando valores elevados de área de contribuição, associadas às menores declividades dessa região, o que acarretou em maiores valores do índice de umidade topográfica.

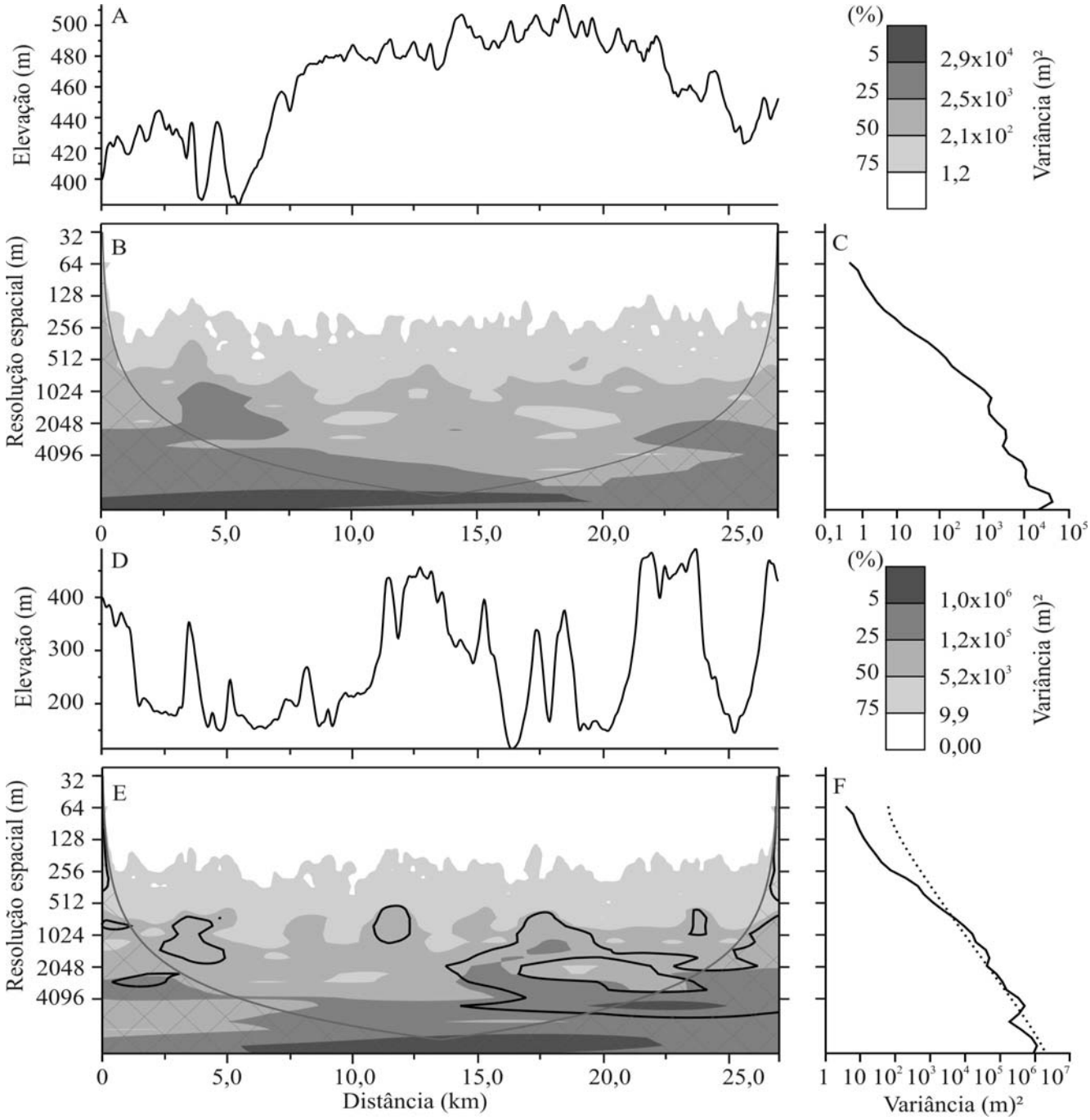
Transformada wavelet

Na região do Planalto, o atributo elevação possui um comportamento que, conforme aumenta a resolução espacial, também aumenta a variância, e a distribuição dos tons de cinza ocorre de maneira quase que horizontal ao longo de todo o transeto (Figura 2 B). Há uma

maior variabilidade entre 4 e 5 km do início do transeto, para valores de resolução de 900 a 2100 m, indicando a presença de pequenos talvegues e mudanças abruptas da elevação (Figura 2 A). Esse é o padrão típico de um evento local, não havendo um ciclo repetitivo desse padrão no restante do isograma e, portanto, não se pode atribuir uma resolução a este atributo (Yates et al., 2006). O escalograma da Figura 2 C indica uma variância crescente com o aumento da resolução, embora os valores não sejam estatisticamente significativos.

Três são os locais de variância significativa para os dados de elevação no transeto localizado no Rebordo (Figura 2 E). O primeiro localizado entre 2,5 e 5 km do início do transeto na faixa de resoluções de 700 a 1500 m; o segundo na região de 11,5 km para resoluções de 512 a 1024 m; e o terceiro iniciando em 13,5 km, estendendo-se até o final do transeto, o que mostra um grande padrão de variabilidade. Contudo, nesse último, existem duas resoluções espaciais predominantes, uma em torno de 1024 m e outra localizada próxima aos 4096 m. Todos esses eventos estão relacionados à grande variabilidade do atributo elevação demonstrado na Figura 2 D. Nas distâncias em que ocorreram os maiores valores de variância significativa, também estão localizadas mudanças abruptas no relevo.

As informações do isograma da Figura 2 E não revelam a presença de um padrão cíclico na variabilidade. Nesse tipo de gráfico, resoluções associadas a eventos cíclicos ao longo de todo o transeto são o foco deste estudo. Entretanto, a análise de significância para a variância global (Figura 2 F) demonstra que existem resoluções associadas a padrões de variabilidade entre 550 e 4200 m. Ressalta-se que essas informações devam ser tomadas com cautela, uma vez que não ocorreram eventos cíclicos no isograma, e os dados de variância global sejam apenas levemente superiores aos do espectro de ruído vermelho. Também é importante frisar que áreas contidas sob o cone de influência não devam ser consideradas devido a efeitos de bordas nos dados (Torrence & Compo, 1998). Observa-se que uma grande área de variância significativa encontra-se na região na Figura 2 E.



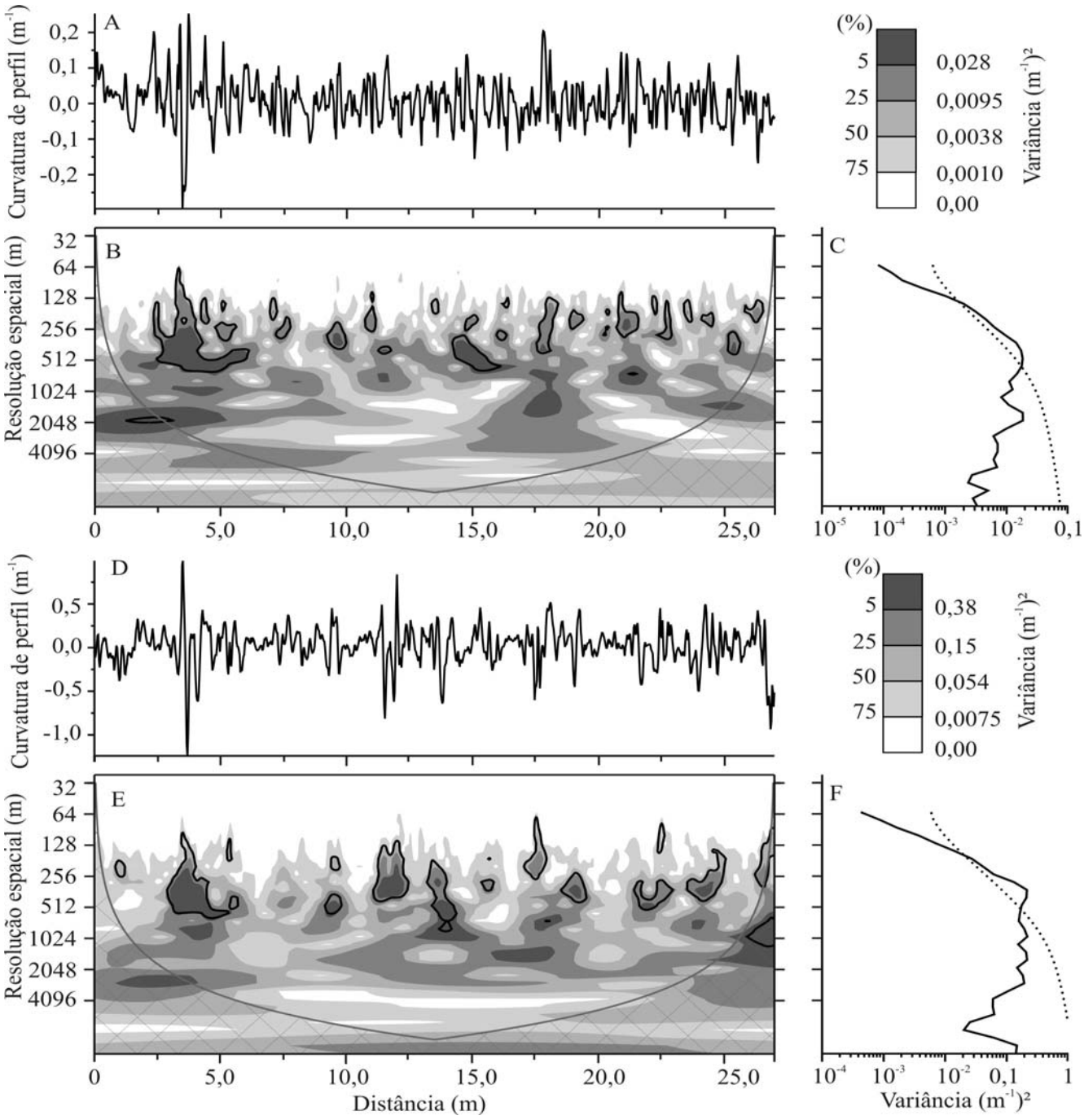
258
259

260 **Figura 2.** Elevação no Planalto (A) e Rebordo (D). Isogramas de variância wavelet no
261 Planalto (B) e Rebordo (E). Regiões quadriculadas representam o cone de influência.
262 Escalogramas da variância total no Planalto (C) e Rebordo (F). Análise de significância foi
263 realizada com o ruído vermelho como espectro de fundo ao nível de 5% de significância nos
264 isogramas (isolinhas) e escalogramas (linha tracejada). A legenda dos níveis de cinza indica o
265 valor e a quantidade de variância relativa em cada nível acima no isograma.

A distribuição da variância do atributo curvatura em perfil apresentou um padrão cíclico ao longo dos 27 km do transeto para a região do Planalto (Figura 3 B). Entre 2,5 e 5 km existe uma grande variabilidade nos valores desse atributo (Figura 3 A), o que foi revelado pela transformada wavelet. Para a sequência do transeto, o que se verifica é que as regiões de variabilidade significativa se repetem praticamente em intervalos regulares. Grande parte desse padrão está situado entre as resoluções de 128 e 512 m, que corresponde à variância total significativa na Figura 3 C. Segundo Yates et al. (2006), nos casos em que há uma região cíclica e contígua no padrão da variabilidade, a opção pela resolução espacial deve ser feita ao centro desse padrão. Adotando esse critério, uma resolução importante para descrever o padrão espacial da curvatura em perfil na região do Planalto seria em torno de 256 m.

O padrão da variância da transformada wavelet para a curvatura em perfil no transeto localizado no Rebordo também apresentou ciclos de ocorrência repetitivos (Figura 3 E). Para esse atributo é possível perceber que as variações de curvatura envolveram amplitudes maiores, localizadas em 3; 12; 14; 17,5; 22,5; 24 e 27 km (Figura 3 D), do que aquelas percebidas na área do Planalto (Figura 3 A), o que provocou um aumento no valor da resolução espacial (variância total com pico em torno de 380 m – Figura 3 F), confirmando o aumento da resolução para a curvatura em perfil em áreas do Rebordo.

Duas regiões de resoluções espaciais significativas podem ser verificadas na transformada wavelet dos dados de declividade do transeto localizado na Depressão (Figura 4 B). Uma primeira, com um padrão mais cíclico, situada entre 70 e 512 m, e outra, essa não cíclica, entre as resoluções de 512 e 1100 m, aproximadamente. A observação do espectrograma da variância total (Figura 4 C) indica que a primeira região tem um pico de variância situado por volta de 256 m, o que indica ser essa uma resolução importante para a variabilidade da declividade nessa área. Em estudo realizado por Si & Farrell (2004), foi verificado que o atributo comprimento de rampa possuía uma resolução significativa em 180 m.



291

292 **Figura 3.** Curvatura em perfil no Planalto (A) e Rebordo (D). Isogramas de variância wavelet
293 no Planalto (B) e Rebordo (E). Regiões quadriculadas representam o cone de influência.
294 Escalogramas da variância total no Planalto (C) e Rebordo (F). Análise de significância foi
295 realizada com o ruído vermelho como espectro de fundo ao nível de 5% de significância nos
296 isogramas (isolinhas) e escalogramas (linha tracejada). A legenda dos níveis de cinza indica o
297 valor e a quantidade de variância relativa em cada nível acima no isograma.

298 A avaliação do somatório de todas as áreas tidas como significativas no isograma da
299 Figura 4 E indica ser esse o maior valor entre os gráficos analisados. Tal fato decorre das
300 grandes amplitudes de declividade presentes sistematicamente ao longo de todo o transeto do
301 Rebordo do Planalto. Há um padrão cíclico entre as resoluções de 128 a 1024 e outro entre as
302 resoluções de 1024 a 2048 m, embora esse último não se repita sistematicamente ao longo dos
303 27 km. O gráfico da variância total indica que há um pico na região situada em torno da
304 resolução de 380 m (Figura 4 F). Esse mesmo valor foi relatado como significativo para a
305 determinação do padrão espacial do atributo curvatura em perfil na região do Rebordo (Figura
306 3 E).
307 Como a região do Rebordo do Planalto serve como área de contribuição da drenagem para
308 a Depressão Central, os valores do índice de umidade topográfica foram maiores em áreas do
309 transeto localizado na Depressão do que aquele transeto localizado em área de Rebordo
310 (Figura 5 A e 5 D). Ambos os isogramas apresentam um padrão cíclico da variância
311 significativa ao longo dos 27 km (Figura 5 B e 5 E). A observação dos espectrogramas para
312 essas regiões indica que há uma faixa de resoluções, entre 200 e 800 m, responsável por
313 grande parte da variabilidade nesses transetos (Figura 5 C e 5 F). Se utilizarmos o critério de
314 Yates et al. (2006) e adotarmos a resolução espacial central entre esses dois valores como
315 sendo aquela determinante para explicar o padrão de distribuição espacial do índice de
316 umidade topográfica, esse valor seria de 400 m.
317 A análise feita nos isogramas indicou que três são as resoluções importantes, a saber: para
318 os atributos declividade e curvatura em perfil, tem-se o valor de 256 m em áreas da Depressão
319 e Planalto e de 380 m para áreas do Rebordo; o atributo índice de umidade topográfica teria
320 seu padrão de variabilidade espacial melhor representado em resoluções de 400 m nas áreas
321 da Depressão e no Rebordo. O critério do valor de um quarto dessas resoluções (He et al.,
322 2007) remete a tamanhos de pixels de 64, 95 e 100 m, respectivamente. Quanto ao atributo

elevação não foi possível precisar com segurança a resolução espacial que melhor representa sua variabilidade, contudo, os resultados apontam que valores maiores do que 550 m poderiam ser utilizados.

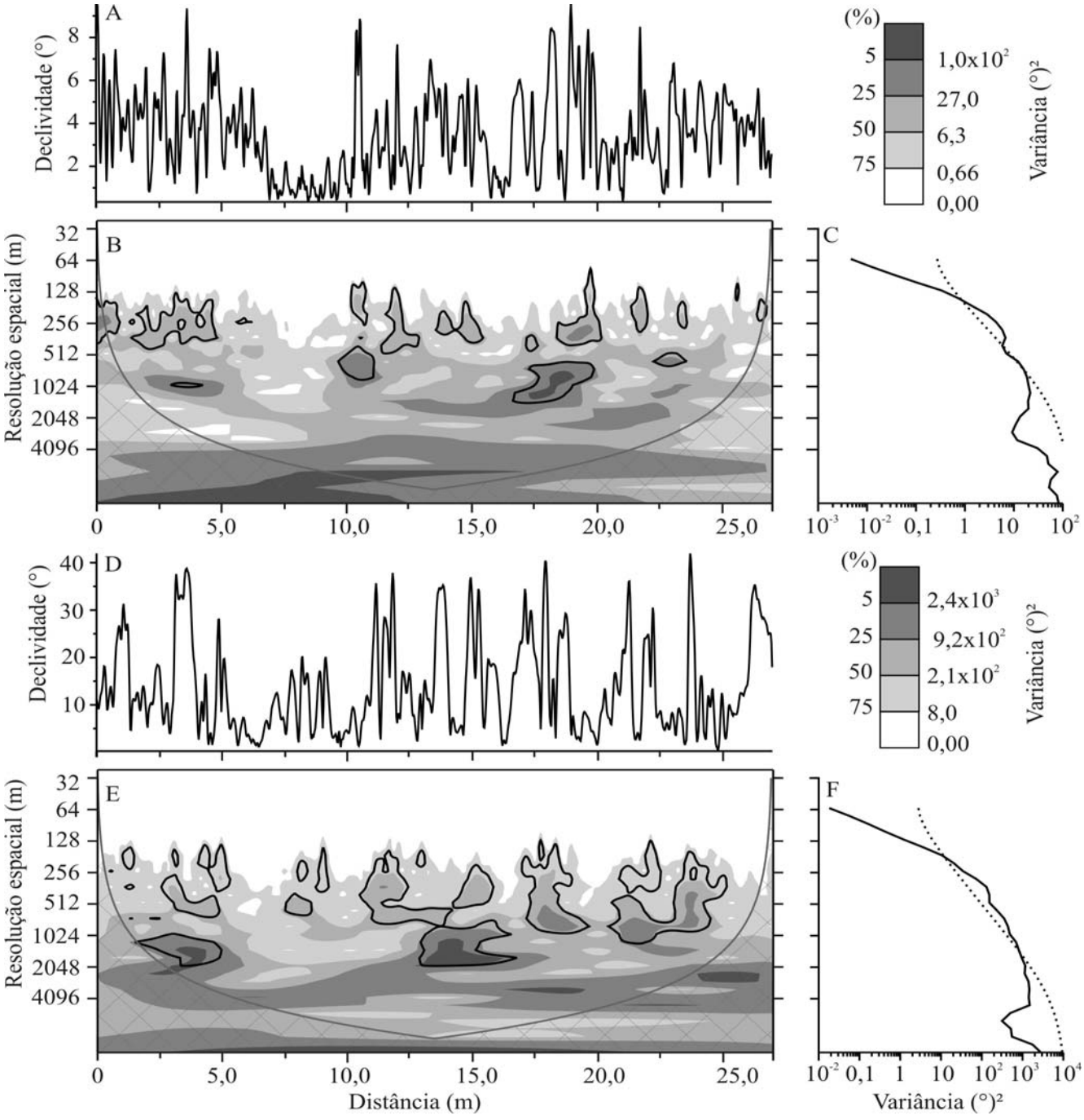


Figura 4. Declividade na Depressão (A) e Rebordo (D). Isogramas de variância wavelet na Depressão (B) e Rebordo (E). Regiões quadriculadas representam o cone de influência. Escalogramas da variância total na Depressão (C) e Rebordo (F). Análise de significância foi

realizada com o ruído vermelho como espectro de fundo ao nível de 5% de significância nos isogramas (isolinhas) e escalogramas (linha tracejada). A legenda dos níveis de cinza indica o valor e a quantidade de variância relativa em cada nível acima no isograma.

A definição da resolução espacial e do tamanho de pixel desses atributos de terreno pode ainda ser utilizada para definir a densidade de amostragem de outras covariáveis ambientais. Si & Farrell (2004) identificaram um padrão de variabilidade espacial similar entre o índice de umidade topográfica e o rendimento do trigo em áreas de estudo do Canadá para resoluções menores do que 140 m. Para He et al. (2007), uma resolução espacial apropriada, determinada pela transformada wavelet para o estudo de padrões associados a atributos de terreno e índices de área foliar, seria de 20 m de resolução.

Os resultados do nosso estudo indicam que sensores remotos com tamanhos de pixel próximos a 60 e 100 m podem ser utilizados em pesquisas que considerem os atributos de terreno declividade, curvatura em perfil e índice de umidade topográfica, ou ainda fenômenos ambientais correlacionados a esses.

O emprego do MDE Topodata (30 m) (Valeriano & Rossetti, 2010) para MDS, conforme informações obtidas neste trabalho, pode representar um grau de detalhamento acima do necessário, o que deverá ser testado em estudos futuros. A aplicação da informação original, oriunda do MDE SRTM com resolução de 90 m, como satisfatória para MDS também deve ser avaliada, assim como a aplicação da transformada wavelet para definição da melhor resolução espacial das plataformas Landsat e Cbers aplicadas ao MDS.

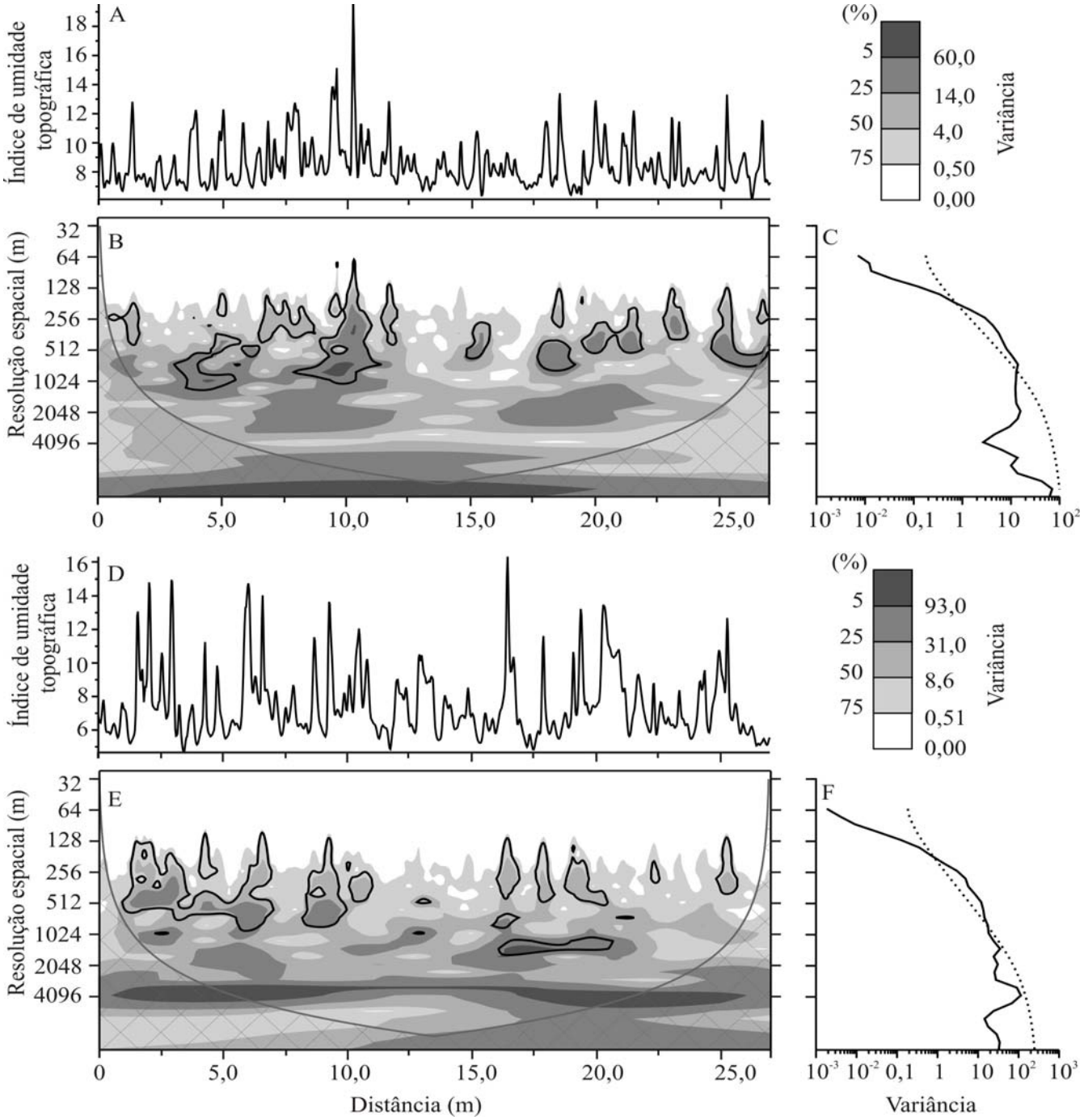


Figura 5. Índice de umidade topográfica na Depressão (A) e Rebordo (D). Isogramas de variância wavelet na Depressão (B) e Rebordo (E). Regiões quadriculadas representam o cone de influência. Escalogramas da variância total na Depressão (C) e Rebordo (F). Análise de significância foi realizada com o ruído vermelho como espectro de fundo ao nível de 5% de significância nos isogramas (isolinhas) e escalogramas (linha tracejada). A legenda dos níveis de cinza indica o valor e a quantidade de variância relativa em cada nível acima no isograma.

CONCLUSÕES

1. Os atributos declividade e curvatura em perfil oriundos de relevos mais suaves têm seu padrão de variabilidade associados à resolução de 256 m. Em relevos mais acidentados, a resolução apropriada é de 380 m. Tamanhos dos pixels em torno de 64 e 95 m são indicados para utilização desses atributos na região de estudo.

2. A resolução de 400 m representou o padrão de variabilidade do índice de umidade topográfica e foi independente da conformação do terreno. O tamanho de pixel de 100 m seria adequado para análises em torno desse atributo.

3. Não foi possível estabelecer para o atributo elevação um valor conclusivo e adequado de resolução espacial e tamanho de pixel.

4. A utilização da wavelet do tipo Morlet permitiu visualizar padrões de variabilidade local e global em atributos de terreno.

AGRADECIMENTOS

Ao CNPq pela bolsa de produtividade em pesquisa do segundo autor e aporte financeiro para auxílio deste trabalho.

REFERÊNCIAS

BIAN, L. Multiscale nature of spatial data in scalling up environmental models. In: QUATTROCHI, D.A; GOODCHILD, M.F. (Ed.) **Scale in remote sensing and GIS.** CRC Press, 1997. p. 13-26.

CHO, E.; CHON, T. Application of wavelet analysis to ecological data. **Ecological Informatics**, v.1, n.3, p.229-233, 2006.

DONG, X.; NYREN, P.; PATTON, B.; NYREN, A.; RICHARDSON, J.; MARESCA, T. Wavelets for agriculture and biology: a tutorial with applications and outlook. **BioScience**, v.58, n.5, p.445-453, 2008.

HE, Y.; GUO, X.; SI, B.C. Detecting grassland spatial variation by a wavelet approach. **International Journal of Remote Sensing**, v.28, n.7, p.1527-1545, 2007.

HENGL, T. Finding the right pixel size. **Computers & Geosciences**, v.32, p.1283-1298, 2006.

JUDEX, M.; THAMM, H.; MENZ, G. Improving land-cover classification with a knowledge based approach and ancillary data. In: Workshop of the EARSeL SIG on Land Use and Land Cover. Braun, M. (Ed.), 2nd., 2006, Bonn. **Proceedings**. Bonn: Center for Remote Sensing of Land Surfaces, 2006. p.184-191.

LABAT, D. Recent advances in wavelet analyses: Part 1. A review of concepts. **Journal of Hydrology**, v.314, p.275-288, 2005.

LARK, R.M; WEBSTER, R. Analysis and elucidation of soil variation using wavelets. **European Journal of Soil Science**, v. 50, p. 185-206, 1999.

MI, X.; REN, H.; OUYANG, Z.; WEI, W.; MA, K. The use of the Mexican Hat and the Morlet wavelets for detection of ecological patterns. **Plant Ecology**, v.179, p.1-9, 2005.

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

SARTORI, P.L.P. Geologia e geomorfologia de Santa Maria. **Ciência & Ambiente**, v.38, p.17-42, 2009.

SCULL, P.; FRANKLIN, J.; CHADWICK, O. A.; MCARTHUR, D. Predictive soil mapping: a review. **Progress in Physical Geography**, v.27, p.171-197, 2003.

SI, B.C. Spatial scaling analyses of soil physical properties: A review of spectral and wavelet methods. **Vadose Zone Journal**, v.7, n.2, p.547-562, 2008.

SI, B.C.; FARRELL R.E. Scale-dependent relationship between wheat yield and topographic indices: A wavelet approach. **Soil Science Society of America Journal**, 2004, v.68, n.2, p.577-587, 2004.

TAGIL, S.; JENNESS, J. GIS-based automated landform classification and topographic, landcover and geologic attributes of landforms around the Yazoren Polje, Turkey. **Journal of Applied Sciences**, v.8, n.6, p.910-921, 2008.

TORRENCE C.; COMPO, G.P. A practical guide to wavelet analysis. **Bulletin of the American Meteorological Society**, v.79, n.1, p.61-78, 1998.

UBERTI, A.A.; KLAMT, E. Relações solo-superfícies geomórficas na encosta inferior do nordeste do Rio Grande do Sul. **Revista Brasileira de Ciência do Solo**, v.8, p.124-132, 1984.

VALERIANO, M. M. ; ROSSETTI, D. F. Topodata: seleção de coeficientes geoestatísticos para refinamento unificado de dados SRTM. São José dos Campos, SP: NPE: Coordenação de Ensino, Documentação e Programas Especiais (INPE-16701-RPQ/853) (Boletim). 2010. 74p.

WILSON, J.P.; GALLANT, J.C. **Terrain Analysis: Principles and Applications.** New York: John Wiley & Sons, 2000. 479p.

WU, W.; FAN, Y.; WANG, Z.; LIU, H. Assessing effects of digital elevation model resolutions on soil–landscape correlations in a hilly area. **Agriculture, Ecosystems & Environment**, v.126, p.209-216, 2008a.

WU, S.; LIB, J.; HUANG, G. A study on DEM-derived primary topographic attributes for hydrologic applications: Sensitivity to elevation data resolution, **Applied Geography**, v.28, p. 210-223, 2008b.

YATES, T.T.; SI, B.C.; FARRELL, R.E.; PENNOCK, D.J. Wavelet spectra of nitrous oxide emission form hummcky terrain during spring snowmelt. **Soil Science Society American Journal**, v.70, p.1110-1120, 2006.

ARTIGO 3 - Mapeamento digital de solos: estratégia de pré-processamento de dados[1]

RESUMO: Mapas de solos têm na borda dos polígonos a região de maior variabilidade, o que leva os pedólogos a divergirem quanto ao delineamento das classes de solos nesses locais. O objetivo deste estudo foi propor uma estratégia de pré-processamento de dados aplicada ao mapeamento digital de solos. Polígonos de solos em um mapa de treinamento foram deslocados para seu interior em 100 e 160 m. Essa estratégia possibilitou que covariáveis localizadas próximas à borda das classes de solos não fossem utilizadas para a geração dos modelos de Árvore Decisão (AD). Três AD, geradas a partir de oito covariáveis preditoras, ligadas aos fatores relevo e organismos, amostradas por um mapa de solos completo e com polígonos deslocados em 100 e 160 m, foram utilizadas para predizer classes de solos. O modelo de AD a partir de observações distantes 160 m da borda dos polígonos no mapa original é menos complexo e tem melhor desempenho preditivo.

Termos de indexação: mapa cloroplético, pedometria, levantamento de solos.

Digital soil mapping: strategy for data preprocessing

SUMMARY: Soil maps have at its polygons edge the region of highest variability, which leads pedologists to disagree about the proper delineation of soil classes at those locations. The aim of this study was to propose a preprocessing strategy applied to digital soil mapping. Soil polygons on a training map were displaced in its inward direction by 100 and 160 m. This strategy has enabled that data covariates located near at the border of soil classes were not used for Decision Tree (DT) model adjusting. Three DT models derived from eight predictors covariate, related with factors relief and organisms, sampled by a complete soil map and by polygons displaced 100 and 160 m were used to predict soil classes. The DT model derived from observations distant 160 m of the boundary between polygons in the original map is less complex and has a better predictive performance.

Index terms: choropleth map, pedometrics, soil survey.

INTRODUÇÃO

Mapas de solos podem ser preditos a partir de informações pré-existentes de solos. O modelo 'scorpan' (McBratney et al., 2003) contempla informações existentes sobre classes e propriedades do solo que possam auxiliar na predição e no mapeamento digital de solos (MDS) em áreas onde a informação espacial sobre o solo não existe ou está indisponível na

escala demandada. De acordo com Qi & Zhu (2003), a ideia básica da formalização das relações solo-paisagem contidas em mapas cloropléticos de solos consiste em se reverter o levantamento de solos pela aplicação de técnicas de mineração de dados. As informações contidas nos polígonos de solos em mapas cloropléticos têm demonstrado sua aplicabilidade ao MDS (Crivelenti et al., 2009; Giasson et al., 2011; ten Caten et al., 2011a).

Nos mapas de solos, o delineamento das unidades de mapeamento é feito através da interpretação visual de pares estereoscópicos, em escala compatível conforme o objetivo do mapa (Dalmolin et al. 2004). A posição espacial dos polígonos de solos implica relações entre as diferentes classes de solos e as condições ambientais presentes na paisagem. Ao delinear as bordas dos polígonos, o pedólogo é conduzido pelo seu conhecimento tácito das relações entre os múltiplos planos de informação que impactam na gênese do solo local, como a geologia, o relevo e o uso da terra (Qi & Zhu, 2003).

Intrínseco ao próprio método está presente a subjetividade. Variações tênues ou graduais nas condições ambientais são de difícil localização através desse método (Zhu et al., 2001), o que gera uma incerteza quanto à real localização na paisagem das transições entre classes de solos. A utilização desses polígonos como referência para o treinamento de modelos preditivos implicará a adição de informações desviadas. Essas, por sua vez, terão diferentes impactos na qualidade preditiva dos modelos (Qi, 2004).

Na busca por melhorar a qualidade do banco de dados utilizado para predizer classes de solos, Qi & Zhu (2003) utilizaram apenas as informações que se posicionassem próximas da moda de cada covariável preditora. Através da construção de um histograma para cada covariável, os autores descartaram aqueles dados que estavam fora da moda do conjunto de dados. A acurácia dos modelos desenvolvidos por árvore de decisão no conjunto de dados filtrados chegou a alcançar 86%, ao passo que modelos utilizando o conjunto total de dados obtiveram uma média de 75%. Para esses autores, a metodologia de filtragem foi efetiva como estratégia de pré-tratamento de dados aplicados ao MDS. Contudo, Schmidt et al., (2008) advertem que essa metodologia por histograma tem como desvantagem a necessidade de se construir um histograma para cada covariável preditora, gerando dificuldade em estudos com número elevado de covariáveis e observações.

Métodos para a seleção das observações são utilizados quando uma grande quantidade de amostras está disponível. Estratégias apropriadas para a seleção das observações conduzem a melhores resultados do que quando os modelos são ajustados ao total de observações disponíveis (Schmidt et al., 2008). O desafio está em fazer mais com menos (Liu & Motoda, 2002), ou seja, extrair um conjunto de dados representativo dos dados originais, utilizando um

conjunto menor de dados, facilmente manipulável pelos algoritmos. Nesse caso, há ganho de acurácia e velocidade de processamento dos dados.

Estudos em MDS buscam avaliar a qualidade das covariáveis e das observações, visando potencializar a acurácia das predições (Schmidt et al., 2008). Entre os estudos que consideram as observações utilizadas no MDS estão os que buscam definir a melhor densidade de amostragem (Moran & Bui, 2002; Scull et al., 2005; Grinand et al., 2008; Schmidt et al., 2008), embora sejam raros os estudos que se preocupam com o padrão das observações dentro de cada classe de solos a ser predita (Qi & Zhu, 2003; Qi, 2004).

Avaliando a aplicação de regressões logísticas múltiplas para a predição de classes de solos, ten Caten et al. (2011b) observaram a confusão dos modelos entre classes próximas na paisagem. Observou-se que os maiores percentuais de confusão da predição ocorreram entre as quatro distintas subordens de Argissolos, uma vez que essas classes de solo ocupam posições muito semelhantes na paisagem. Para os autores, essa dificuldade por parte dos modelos pode ter origem no próprio delineamento do mapa original que serviu de treinamento, já que o solo não tem uma transição abrupta, como as classes de solo no mapa original (com polígonos cloropléticos) ou, ainda, devido a diferenças muito tênues entre os atributos do terreno (covariáveis ambientais), que podem não apresentar nenhum tipo de gradiente na borda dos polígonos das classes de solo.

A influência das áreas de transição entre diferentes classes de solos também foi registrada em estudo realizado por Carvalho et al. (2009). Segundo esses autores, devido à prática da cartografia baseada em polígonos (Booleana) adotada no modelo convencional de mapas de solos, ocorre a tendência, na construção de mapas digitais gerados pela aplicação da metodologia da lógica nebulosa, ao aparecimento de áreas que estariam relacionadas a zonas de transição entre duas ou mais unidades de solo. Isso acarretaria o surgimento de delineamentos inexistentes no mapa convencional utilizado para treinamento dos modelos.

A definição da exata posição das bordas dos polígonos é questão controversa entre os pedólogos. Legros (2005) avaliou a qualidade do delineamento das unidades de mapeamento executadas por 20 diferentes pedólogos em uma mesma área. Foi identificado que os polígonos tendem a se sobreporem. Contudo, pequenos desvios no delineamento são ocasionados pela percepção de cada pedólogo das variações das informações contidas nas fotografias aéreas. Esses desvios do delineamento possibilitam definir uma região de incerteza quanto à mais adequada localização para a borda dos polígonos (Figura 1). Nessa região de incerteza, as covariáveis preditoras podem estar apresentando uma contribuição duvidosa à qualidade dos modelos preditivos.

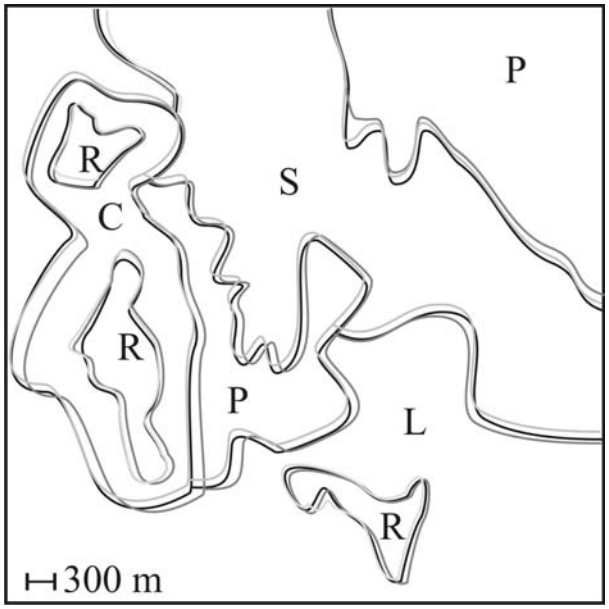


Figura 1. Classes de solos delineadas por três diferentes pedólogos, indicando a região de incerteza na definição do posicionamento da borda dos polígonos. Cambissolo (C), Neossolo (R), Argissolo (P) e Latossolo (L). Adaptado de Legros (2005).

O objetivo deste estudo foi avaliar o impacto de uma estratégia de pré-processamento de dados aplicada ao MDS. Foi verificado o efeito da não utilização de informações derivadas de covariáveis preditoras presentes nas bordas dos polígonos de solos em modelos por árvore de decisão para a predição de classes de solos. Este estudo propõe uma metodologia de amostragem que exclui amostras desviadas e seleciona observações que contém as características das covariáveis preditoras em cada classe de solos.

MATERIAL E MÉTODOS

Mapa de solos para treinamento

Para a realização do presente estudo, foi utilizado como teste o levantamento de solos semidetalhado na escala 1:50.000, realizado por Klamt et al. (2001) para o município de São Pedro do Sul, localizado na região central do Estado do Rio Grande do Sul, com uma extensão de 874 km². As classes de solos ao nível de subordem foram vetorizadas no programa ArcGIS 9.3 (ESRI, 2008).

A proposta deste estudo consiste em reposicionar cada classe de solo presente no mapa original para dentro do polígono originalmente definido pelo pedólogo. Com isso, as regiões de maior incerteza quanto à real posição das bordas dos polígonos de classes de solos não são amostradas para a utilização nos modelos preditivos. Para verificar o efeito da metodologia no

conjunto de dados a serem gerados, foram utilizadas as distâncias de 100 e 160 m em relação à posição original de cada classe de solo (Figura 2).

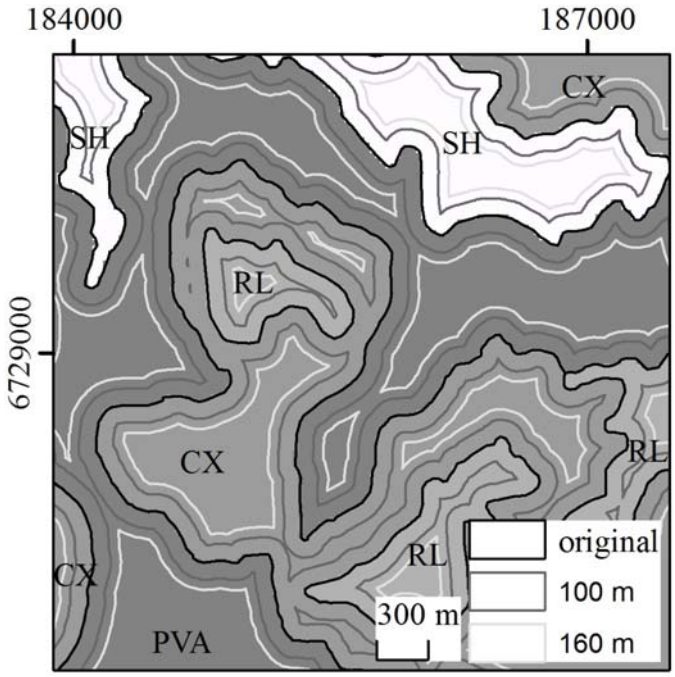


Figura 2. Estratégias de amostragem das covariáveis preditoras adotadas no estudo, usando o mapa de solos original, a 100 m e 160 m da borda dos polígonos. Solos Hidromórficos (SH - Planossolos e Gleissolos), Cambissolo Háplico (CX), Neossolo Litólico (RL) e Argissolo Vermelho-Amarelo (PVA).

Como todos os polígonos de solos são reposicionados na sua direção interna, a faixa de dados efetivamente descartada foi de 200 e 320 m nas bordas de polígonos vizinhos. Os valores de 100 e 160 m foram definidos a partir de uma análise visual dos gradientes de valores ocorrentes nas covariáveis preditoras próximas à borda dos polígonos. Os novos polígonos foram criados a partir da função Buffer do programa ArcGIS 9.3. A metodologia de amostragem proposta não implicou que alguma classe de solo não fosse amostrada em virtude de representar uma diminuição da área total compreendida por cada classe de solo.

Covariáveis preditoras

Neste estudo foram utilizadas covariáveis do modelo 'scorpan' (McBratney et al., 2003), relacionadas aos fatores de formação do solo relevo (*r*) e organismos (*o*). O fator organismo foi representado pela covariável desvio padrão do índice de vegetação por diferença normalizada (NDVI), doravante denominada DEPA. A covariável DEPA foi usada em detrimento ao NDVI pelo fato dessa última apresentar diferentes valores ao longo do ano em função dos diferentes usos agrícolas. Para o cálculo da covariável DEPA, foram utilizadas

imagens obtidas no período de fevereiro de 2004 a janeiro de 2005. Esse período foi selecionado devido à maior disponibilidade de imagens com ausência de nuvens durante aquele ano. Todos os cálculos de NDVI foram realizados conforme Jensen (2009). O preditor DEPA foi gerado a partir de dados de oito distintas datas daquele período, obtidas pela plataforma Landsat 5 sensor TM com resolução espacial de 30 m. Cada data teve seu valor de NDVI calculado individualmente, em seguida foi executado o cálculo do desvio padrão do valor do NDVI entre as oito datas para cada pixel no programa ArcGIS 9.3, na função *Raster Calculator*.

O fator relevo foi representado pelas covariáveis elevação (ELEV), declividade (DECL), índice de umidade topográfica (IUT), capacidade de transporte de sedimento (CTS), curvatura planar (PLAN), curvatura de perfil (PERF) e índice de rugosidade do terreno (IRT). As covariáveis foram geradas de acordo com Wilson & Gallant (2000) a partir de um modelo digital de elevação (MDE). O MDE com resolução espacial de 30 m foi derivado de curvas de nível de cartas topográficas com escala 1:50.000. A interpolação das curvas de nível para o formato matriz ocorreu na ferramenta *topo do raster* do programa ArcGIS 9.3, usando a técnica *spline* (Wahba, 1990). No programa SAGA-GIS (Olaya, 2004), os atributos ELEV, DECL, PLAN, PERF, e IRT foram gerados na ferramenta *standard terrain analysis*, e os atributos IUT e CTS com a ferramenta *grid calculator*.

Esses atributos foram tabulados a partir dos polígonos de solos, gerando três distintos conjuntos de dados para construção dos modelos por árvore de decisão: original, 100 m e 160 m.

Árvore de Decisão (AD)

Nos três conjuntos de dados utilizados para desenvolver os modelos, todas as classes de solos foram amostradas proporcionalmente a sua área. O desenvolvimento das AD foi realizado no programa de mineração de dados WEKA 3.6.3 (Hall et al., 2009). Para o processamento dos dados, foi utilizado o algoritmo J48 que apresentou os melhores resultados em estudo realizado por Giasson et al. (2011). O número mínimo de observações por folha (*minNumObj* - WEKA), para cada conjunto de dados foi determinado após uma análise do percentual de observações erroneamente classificadas. Essa análise foi executada com uma série de valores para o número mínimo de observações em cada um dos três conjuntos de dados. O método de poda que mitigava o erro na árvore gerada também foi selecionado (*reducedErrorPruning = True* - WEKA). Durante a fase de geração da árvore, cada conjunto

de dados foi particionado em 70% para geração do modelo e 30% para validação da árvore (*Percentage split* - WEKA).

Mapa de solos

Após a análise da complexidade das AD geradas e do número de observações erroneamente classificadas, as AD foram implementadas no programa ArcGIS 9.3, com a função *Raster Calculator*. As informações derivadas da árvore foram convertidas na função condicional 'con(teste, verdadeiro, falso)' do programa. Essa função permite que os arquivos raster com as covariáveis ambientais sejam processados de acordo com o conjunto de regras derivadas da AD. Como a escala de publicação pretendida é de 1:50.000, em cada mapa de solos gerado foram extraídas regiões de píxeis isolados com área mínima mapeável menor do que um hectare.

Qualidade dos modelos e mapas

A qualidade dos modelos AD foi avaliada a partir do valor percentual de observações erroneamente classificadas em toda a árvore. Este valor é uma das saídas do programa WEKA após o modelo gerado ser validado no conjunto de 30% dos dados reservados. Para seu cálculo, o programa soma todas as observações erroneamente classificadas e as divide pelo total de observações do conjunto teste, multiplicando o resultado por 100 para gerar o valor percentual (Hall et al., 2009). Os mapas de solos gerados a partir dos três distintos conjuntos de dados foram comparados com a geração do índice kappa. O índice kappa é um indicador utilizado para atestar a qualidade dos mapeamentos preditivos (Giasson et al., 2011). A matriz dos erros para o cálculo do índice kappa foi executada conforme Congalton (1991).

RESULTADOS E DISCUSSÃO

A metodologia de pré-processamento proposta implicou a redução do número total de observações disponíveis para gerar os modelos por árvore de decisão. No conjunto de dados derivados do mapa original, estão presentes 100% das informações utilizadas neste estudo. O conjunto de dados gerado a partir do deslocamento dos polígonos em 100 m reteve 60% dos dados originais. Por ocasião de um deslocamento de 160 m nos polígonos de solos, 43% dos dados originais foram tabulados para gerar a árvore de decisão. Esses percentuais de dados são superiores aos 30% utilizados por Grinand et al. (2008) e aos 25% utilizados por Moran & Bui (2002) para o ajuste de árvores de decisão aplicadas ao mapeamento digital de solo.

Os conjuntos de dados derivados de polígonos de Neossolos Litólicos deslocados em 100 e 160 m têm medidas descritivas distintas do conjunto de dados original (Figura 3). Entre as

observações mais distantes das características centrais de cada covariável preditora estão os outliers. Com exceção do atributo elevação, as demais covariáveis continham outliers entre os dados amostrados a partir do polígono original. Com a aplicação dos deslocamentos nos polígonos de solos, observações próximas das bordas dos polígonos foram descartadas, e a presença de valores distantes em cada covariável não ocorreu. O pré-processamento significou mudanças similares no padrão dos dados originados a partir das demais classes de solos, embora esses não estejam aqui representados.

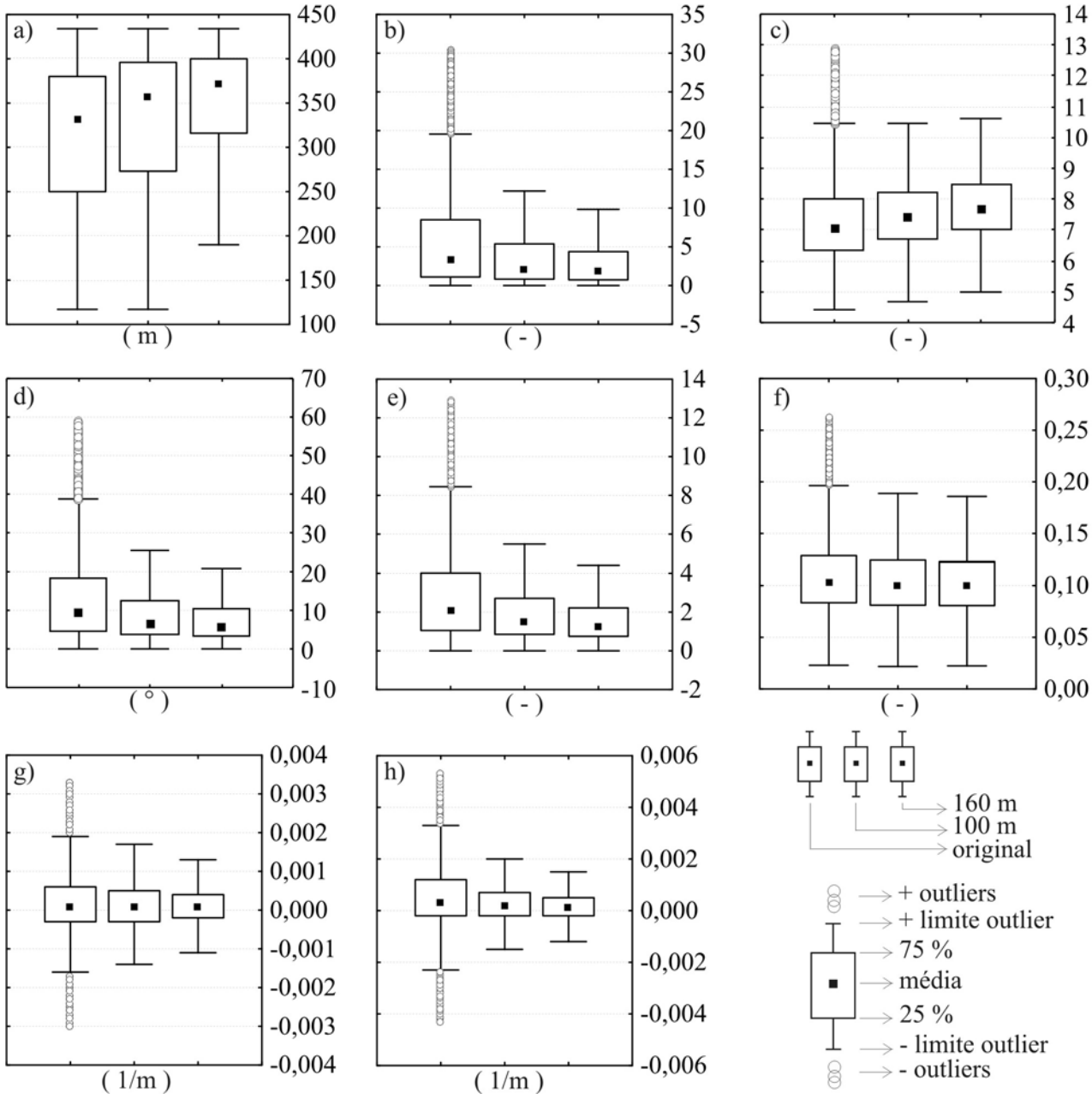


Figura 3. Boxplot das covariáveis preditoras gerados a partir dos dados amostrados pela classe dos Neossolos Litólicos nas três situações de posição das bordas dos polígonos. a) elevação (m), b) capacidade de transporte de sedimento, c) índice de umidade topográfica, d)

declividade (°), e) índice de rugosidade do terreno, f) desvio padrão do NDVI, g) curvatura planar (1/m) e h) curvatura de perfil (1/m). Unidades de cada covariável estão entre parênteses. O limite outlier foi calculado a partir da altura da caixa central do boxplot multiplicado por 1,5. Foram tomados como outliers os valores além do limite outlier.

A análise visual dos gráficos indica que o pré-processamento não alterou a distribuição dos dados. A posição da média em relação ao conjunto de 50% do total de dados indica que esses não têm distribuição normal. Esse padrão se repetiu para os dados oriundos dos deslocamentos de 100 e 160 m. Embora tenha ocorrido uma diminuição da dispersão dos dados, esses passaram a se situar menos distantes da média. Esse comportamento é menos marcante nos dados oriundos da covariável desvio padrão do NDVI (Figura 3f). Isso pode ser atribuído ao fato de que na área de estudo os Neossolos Litólicos estejam situados em áreas de encosta, as quais se encontram cobertas por mata nativa. Da posição dos polígonos no mapa original até os polígonos deslocados 160 m não ocorrem mudanças marcantes do uso da terra no local.

AD obtidas pelos três conjuntos de dados têm um comportamento distinto quanto ao número de observações erroneamente classificadas (Figura 4). As AD geradas a partir do conjunto total de dados obtiveram uma classificação das observações com um erro mínimo de 38,5%. Os demais conjuntos de dados alcançaram erros mínimos de 32 e 28% para deslocamentos de 100 e 160 m, respectivamente. Isso indica que houve um benefício do pré-processamento dos dados para o melhor ajuste do modelo de AD. Contudo, as árvores ajustadas aos dados com menor presença de observações desviadas não classificaram adequadamente cerca de um terço dos dados. Isso pode estar ocorrendo em decorrência da complexidade da distribuição espacial dos solos no local, que não foi representada pelas oito covariáveis preditoras escolhidas para este estudo. Além disso, pode estar ligado a características do mapa de referência como a escala e o delineamento dos polígonos.

Nos três conjuntos de dados, o menor percentual de erros foi alcançado com árvores com um maior número de nós e folhas terminais. O número de observações erroneamente classificadas permaneceu praticamente inalterado na medida em que era permitido ao programa agrupar um maior número de observações nas folhas finais. Contudo, próximo ao número de 30 folhas, os três conjuntos de dados apresentaram uma tendência de aumento das observações erroneamente classificadas. Acredita-se que a partir desse limite a árvore seja demasiadamente simplificada e torna-se incapaz de predizer a complexidade presente nos dados.

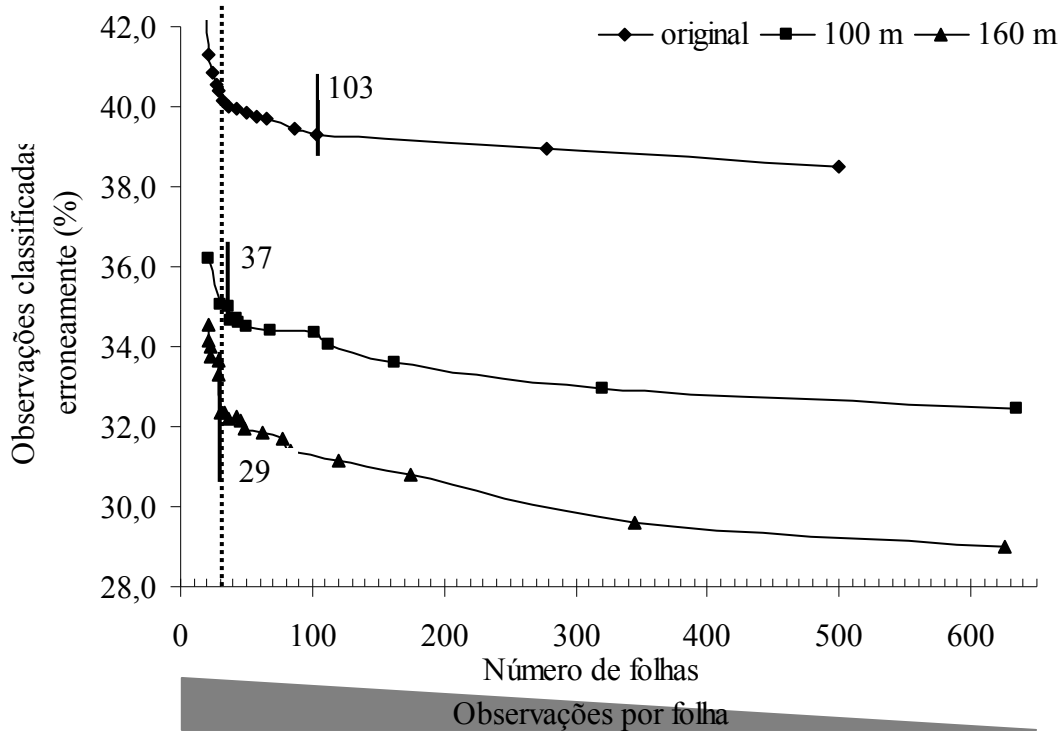


Figura 4. Relação entre o número de folhas e observações classificadas erroneamente por árvore de decisão. Conjuntos de dados sem deslocamento (original) e com deslocamento de 100 m e 160 m da borda dos polígonos. Os números próximos às barras verticais indicam o número mínimo de folhas para que todas as classes de solos fossem preditas com aquele conjunto de dados. A linha tracejada vertical indica o número de 30 folhas como sendo aquele a partir do qual ocorre um súbito incremento do número de observações erroneamente classificadas nos três conjuntos de dados.

AD geradas a partir dos dados afastados das bordas dos polígonos de solos têm mais poder preditivo. Observando o número mínimo de folhas para que as árvores de decisão ainda mapeassem as sete classes de solos presentes na área de estudo, verifica-se que, para os dados oriundos do deslocamento de 100 m, o número mínimo de folhas foi de 37. Quando o deslocamento efetuado foi de 160 m, o número mínimo de folhas foi de 29. Por outro lado, com a utilização da totalidade das observações presentes na área de estudo, somente com árvores de maior complexidade, acima de 103 folhas, foi possível predizer todas as classes de solos. Esses resultados indicam que a presença de observações desviadas demanda árvores mais complexas para capturar o padrão presente entre os dados. Em estudo realizado por Giasson et al. (2011), a partir de 1.333 observações aleatoriamente distribuídas, os autores

verificaram que entre as árvores testadas apenas as mais complexas eram capazes de predizer todas as classes de solos.

Os resultados não permitem inferir qual seria a distância adequada de deslocamento das bordas dos polígonos de solos. Essa distância será uma função de vários fatores, como a escala, a complexidade dos elementos que governam a distribuição do solo, a experiência e habilidade do pedólogo que mapeou a região que está sendo utilizada para gerar os modelos. Possivelmente, em um mesmo mapa, poderiam ser adotadas distâncias distintas, uma vez que as classes de solos apresentam um gradiente de complexidade nas suas transições. No entanto, a análise da figura 4 permite afirmar que o pré-processamento pelo deslocamento dos polígonos de solos, evitando utilizar as observações localizadas nas regiões de maior incerteza do mapa, possibilita AD mais precisas e menos complexas.

Mapas de solos foram gerados por AD com um número mínimo de 30 folhas, a partir dos três conjuntos de dados (Figura 5). Esse número foi adotado em virtude de ser o valor que permite a melhor relação custo benefício entre complexidade e poder preditivo dos modelos como verificado na figura 4. Embora, como anteriormente descrito, esse número mínimo de folhas implique que as árvores geradas a partir do conjunto original e com deslocamento de 100 m não predizem todas as classes de solos.

Visualmente os Solos Hidromórficos (Planossolos e Gleissolos) foram espacializados de maneira similar ao mapa de treinamento. Nos três mapas preditos, a classe Argissolo Bruno-Acinzentado foi espacializada em uma posição intermediária na paisagem entre as áreas de várzea e áreas mais elevadas das colinas, onde estão localizados os Argissolos Vermelho-Amarelos. O modelo gerado pelo conjunto de dados oriundos do deslocamento de 160 m possibilitou a espacialização da classe Argissolo Acinzentado, embora essa classe, presente em outras regiões do mapa original, não esteja localizada na área dos extratos utilizados para demonstrar os resultados (Figura 5).

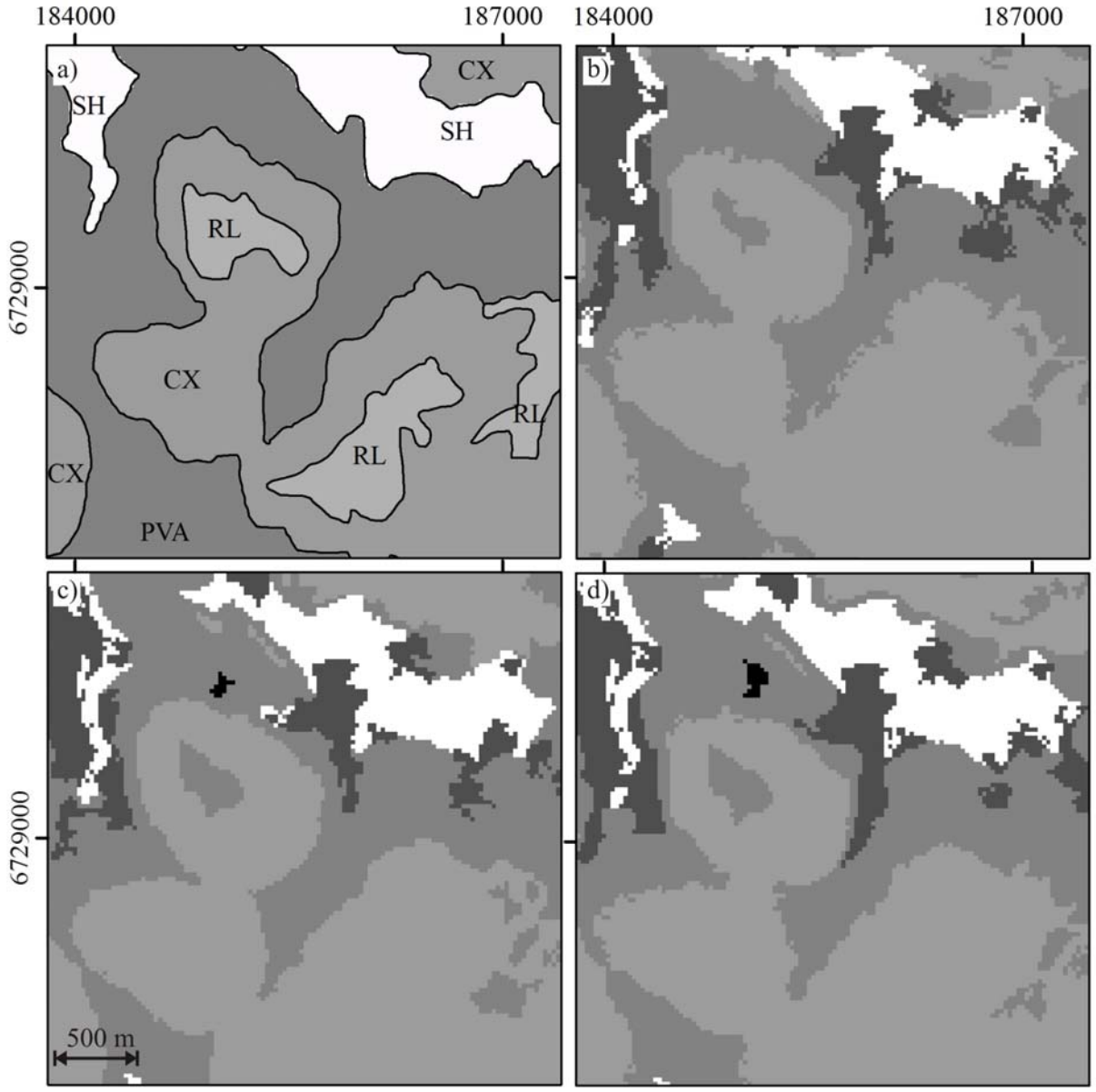


Figura 5: a) Extrato do mapa de solos utilizado para treinamento. b) Predito a partir da totalidade dos dados. c) Predito a partir de deslocamento de 100 m nos polígonos. d) Predito a partir de deslocamento de 160 m nos polígonos. Solos Hidromórficos (SH), Cambissolo Háplico (CX), Neossolo Litólico (RL) e Argissolo Vermelho-Amarelo (PVA).

A comparação entre os três mapas de solos gerados e o mapa original permite verificar um satisfatório grau de concordância entre esses (Quadro 1). Aproximadamente, em 60% dos pontos (pixels) sobre a paisagem, a mesma classe existente no mapa original foi atribuída pelos modelos aos mapas preditos. Esse valor é superior aos valores encontrados na literatura de 43,0% (Crivelenti et al., 2009), 38,6% (Coelho & Giasson, 2010) e 51,8% (Giasson et al., 2011). Apesar de uma qualidade de predição superior obtida neste trabalho, em relação aos

311 trabalhos encontrados na literatura, observa-se que, com o deslocamento de 100 m, a classe
312 dos Argissolos Acinzentados não foi predita. Esse fato provavelmente decorre de que a
313 limitação de um número máximo de 30 folhas seja uma restrição muito forte à complexidade
314 das informações existentes entre os dados advindos de posições na paisagem onde ocorram os
315 Argissolos Acinzentados. Nesse caso, há a necessidade de que, para os conjuntos de dados
316 mais dispersos, árvores mais complexas sejam geradas para a predição de todas as classes de
317 solos.
318 Quadro 1. Índice kappa indicando o grau de concordância entre os mapas gerados a partir dos
319 três conjuntos de dados deste estudo e o mapa de treinamento.

	Mapa de treinamento	Original*	100 m*
Original*	60,64	-	-
100 m*	60,16	90,5	-
160 m	59,34	87,6	88,85

320 * Classe de solos dos Argissolos Acinzentados não foi predita.
321
322 A comparação entre os três mapas de solos indica um elevado grau de similaridade entre
323 esses (Quadro 1). Os três mapas foram idênticos em torno de 90% dos pontos na
324 paisagem,embora tenha de se considerar que uma das classes de solos não tenha sido mapeada
325 pela AD gerada por dois dos conjuntos de dados. Se considerarmos que o pré-processamento
326 dos dados implicou uma redução do volume de observações, o conjunto de dados derivado do
327 deslocamento de 160 m na borda dos polígonos, proposta de se fazer mais com menos (Liu &
328 Motoda, 2002), foi alcançada. O índice kappa de 59,34% em relação ao mapa original foi
329 alcançado com a totalidade das classes mapeadas, e com um volume de dados a ser
330 manipulado, aproximadamente 60% menor do que o volume original.
331 O pré-processamento de dados apresentado difere da metodologia empregada por Qi &
332 Zhu (2003), a qual emprega o conceito do histograma e da moda em dois aspectos principais.
333 O primeiro deles está ligado ao fato de que a seleção das observações a partir de sua relação
334 com a moda de uma covariável poderá representar a exclusão de informações relevantes.
335 Podem ocorrer situações em que uma covariável apresente características multimodais em
336 uma mesma classe de solo. Assim, é difícil determinar quantas são e quais são as modas
337 presentes nos dados. Outro aspecto diz respeito ao volume de processamento necessário. O
338 deslocamento dos polígonos de solos é automatizado, ao passo que a construção e análise dos
339 histogramas irão demandar mais tempo para a análise.

CONCLUSÕES

1. O pré-processamento das observações reduz o volume de dados a ser manipulado no MDS.
2. Observações presentes na borda dos polígonos de solos aumentam o número de observações erroneamente classificadas por árvore de decisão.
3. Árvores de decisão geradas a partir de observações afastadas das bordas dos polígonos de solos são menos complexas e têm maior poder preditivo em MDS.
4. Mapa de solos obtido por árvores de decisão a partir de observações não desviadas tem mais similaridade ao mapa de treinamento.
5. A metodologia apresentada é de simples implementação e fácil aplicação para MDS.

AGRADECIMENTOS

À CAPES pelos recursos utilizados na execução do trabalho e ao CNPq pela bolsa de produtividade em pesquisa do segundo autor e aporte financeiro para auxílio desse trabalho.

LITERATURA CITADA

CARVALHO, C.C.N.; FRANCA-ROCHA, W. & UCHA, J.M. Mapa digital de solos: Uma proposta metodológica usando inferência fuzzy. Rev. Bras. Eng. Agríc. Ambient., 13:46-55, 2009.

CRIVELENTI, R. C.; COELHO, R. M.; ADAMI, S. F. & OLIVEIRA, S. R. de M. Mineração de dados para a inferência de relações solo-paisagem em mapeamentos digitais de solo. Rev. Agro. Bras. 44:1707-1715, 2009.

CONGALTON, R.G. A review of assessing the accuracy of classification of remotely sensed data. Remote Sens. Environ. 37: 35-46, 1991.

COELHO, F.F. & GIASSON, E. Comparação de métodos para mapeamento digital de solos com utilização de sistema de informação geográfica. Cienc. Rural, 40:2099-2106, 2010.

DALMOLIN, R. S. D.; KLAMT, E.; PEDRON, F. de A.; AZEVEDO, A.C. de. Relação entre as características e o uso das informações de levantamentos de solos de diferentes escalas. Cienc. Rural, 34:1479-1486, 2004.

ESRI Environmental Systems Research Institute, Inc., Redlands, CA, 2008.

GIASSON, E.; SARMENTO, E.C.; WEBER, E.; FLORES, C.A. & HASENACK, H. Decision trees for digital soil mapping on subtropical basaltic steeplands. Sci. Agrí., 68:167-174, 2011.

374 GRINAND, C.; ARROUAYS, D.; LAROCHE, B. & MARTIN, M. Extrapolating regional
375 soil landscapes from an existing soil map: Sampling intensity, validation procedures, and
376 integration of spatial context. Geoderma., 143:180-190, 2008.
377 HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P. & WITTEN,
378 I.H. The WEKA Data Mining Software: An Update. SIGKDD Explorations Newsletter,
379 11:10-18, 2009.
380 JENSEN, J.R. Sensoriamento Remoto do Ambiente: uma perspectiva em recursos terrestres.
381 São José dos Campos. Ed. Parêntese. 2009. 598p.
382 KLAMT, E.; FLORES, C. A. & CABRAL, D. R. Solos do Município de São Pedro do Sul.
383 Departamento. de Solos/CCR/UFSM. Santa Maria, 2001, 96p.
384 LEGROS, J. P. Mapping of the soil. Enfield: Science Publisher, 2005, 411p.
385 LIU, H. & MOTODA, H. On Issues of Instance Selection, Data Min. Knowl. Disc., 6:115-
386 130, 2002.
387 MCBRATNEY, A. B.; MENDONCA SANTOS, M. L. & MINASNY, B. On digital soil
388 mapping. Geoderma, 117:3-52, 2003.
389 MORAN, C.J. & BUI, E.N. Spatial data mining for enhanced soil map modeling. Int. J.
390 Geogr. Inf. Sci., 16:533-549, 2002.
391 OLAYA, V. A gentle introduction to SAGA GIS. The SAGA user group, Gottingen,
392 Germany, 208p.
393 QI, F. Knowledge discovery from area-class resource maps: data preprocessing for noise
394 reduction. Transactions in GIS. 8:297–308, 2004.
395 QI, F. & ZHU, A. X. Knowledge discovery from soil maps using inductive learning. Int. J.
396 Geogr. Inf. Sci., 17:771-795, 2003.
397 TEN CATEN, A.; DALMOLIN, R.S.D.; PEDRON, F. de A. & MENDONÇA-SANTOS, M
398 de L. Extrapolação das relações solo-paisagem a partir de uma área de referência. Cienc.
399 Rural., 41:812-816, 2011a.
400 TEN CATEN, A.; DALMOLIN, R.S.D.; PEDRON, F. de A. & MENDONÇA-SANTOS, M
401 de L. Regressões logísticas múltiplas: fatores que influenciam sua aplicação na predição de
402 classes de solos. R. Bras. Ci. Solo, 35:53-62, 2011b.
403 SCHMIDT, K; BEHRENS, T. & SCHOLTEN, T. Instance selection and classification tree
404 analysis for large spatial datasets in digital soil mapping. Geoderma, 146:138-146, 2008.
405 SCULL, P.; FRANKLIN, J.; CHADWICK, O.A. & MCARTHUR, D. Predictive soil
406 mapping: a review. Prog. Phys. Geog., 27:171-197, 2003.

ZHU, A.X.; HUDSON, B.; BURT, J.; LUBICH, K. & SIMONSON, D. Soil mapping using GIS, expert knowledge, and fuzzy logic. Soil Sci. Soc. of Am. J., 65:1463-1472, 2001.

WAHBA, G. Spline models for Observational data. In: CBMS-NSF Regional Conference Series in Applied Mathematics. Philadelphia: Soc. Ind. Appl. Maths., v.59, 1990, 169p.

WILSON, J. P. & GALLANT, J. C. Digital terrain analysis. In:_____ ed. Terrain analysis: principles and applications. New York: Wiley & Sons, 2000, p. 1-27.

ARTIGO 4 - Mapeamento de classe de solos por árvore de decisão: impacto do volume de dados.[1]

Resumo: Informações digitais possibilitam ao Mapeamento Digital de Solos (MDS) um elevado grau de redundância para o ajuste de modelos preditivos de classes e propriedades de solo. Entre esses modelos, a técnica de Árvores de Decisão (AD) tem aplicação crescente devido a sua robustez no tratamento de grandes volumes de dados. Esse trabalho tem como objetivo avaliar o impacto do volume de dados utilizados para gerar os modelos por AD na qualidade dos mapas de solos preditos. A área de estudo, com 889,33 km², está localizada na região do Planalto Médio do Rio Grande do Sul. As relações solo-paisagem foram obtidas a partir de reambulação da área de estudo e delineamento das unidades de mapeamento em cartas topográficas de escala 1:50.000. Seis covariáveis preditoras ligadas aos fatores de formação do solo relevo e organismos, juntamente com os conjuntos de dados de um, três, cinco, 10, 15, 20 e 25% do volume total de dados foram utilizados para gerar os modelos preditivos por AD no programa WEKA. Nesse estudo, densidades de amostragem menores do que 5% implicam modelos preditivos com menor poder de capturar a complexidade da distribuição espacial do solo. Amostragens entre cinco e 15% provocaram melhor relação custo-benefício quanto ao volume de dados a ser manipulado e à capacidade preditiva dos modelos gerados. Dados coletados a campo indicaram uma acurácia dos mapas preditos próxima a 70% para os modelos oriundos dessas densidades de amostragem.

INTRODUÇÃO

O avanço tecnológico em áreas como sensoriamento remoto, velocidade de processamento dos computadores, possibilidade de manipular grandes volumes de dados, métodos quantitativos para descrever padrões espaciais e visualização tridimensional tem possibilitado uma nova oportunidade para predizer processos e propriedades do solo

(Grundwald, 2009). Essa oportunidade foi denominada por McBratney et al. (2003) de Mapeamento Digital de Solos (MDS).

A aplicação de informações digitais ao MDS possibilita um elevado grau de redundância para o ajuste de modelos. Por outro lado, o volume de informações disponíveis irá requerer a manipulação e o processamento de grandes volumes de dados (McBratney et al., 2003). A busca pela melhor relação custo-beneficio entre a densidade de amostragem no banco de dados original e a acurácia preditiva dos modelos tem motivado pesquisas neste sentido.

Entre as metodologias de mineração de dados aplicadas ao MDS, a técnica de Árvores de Decisão (AD) tem aplicação crescente devido a sua robustez no tratamento de grandes volumes de dados (Witten e Frank, 2005). A abordagem por AD tem sido empregada por apresentar a vantagem de possibilitar a expressão das relações solo-paisagem de maneira explícita (Kheir et al., 2010a). As AD foram empregadas em estudos relacionados à erosão superficial e subterrânea de solos (Geissen et al., 2007), à predição de classes de solos (Giasson et al., 2011) e à espacialização de propriedades do solo (Lemercier et al., 2011).

Além de possibilitar o agrupamento e a busca por padrões, a AD possibilita o entendimento de como estes dados são inter-relacionados (Kheir et al., 2010b). A AD não necessita que seja especificada a forma do modelo que será ajustado aos dados, bem como, à medida em que um modelo por AD é construído, ele pode ser convertido em algoritmos que podem ser facilmente implementados por linguagens de programação (Kheir et al., 2010a).

Modelos por AD foram utilizados por Scull et al. (2005) com 39.877 amostras para a predição de classes de solos, em uma área de deserto de 2.590 km², correspondendo a, aproximadamente, um por cento do total de dados disponíveis. Em estudo realizado por Giasson et al. (2011), foram utilizadas 1.333 amostras para a predição de classes de solos em uma área aproximadamente 6,1 km². A densidade de amostras correspondeu a menos de um

por cento do volume de dados original. Qi e Zhu (2003) amostraram, em uma área de 4 km², cerca de 12% dos dados originais para o ajuste de modelos por AD.

Frente à disponibilização de bancos de dados com um crescente número de observações e a diversidade de densidade de amostragens relatada nos estudos realizados, este trabalho tem como objetivo avaliar a influência da proporção do conjunto de dados originais na qualidade do mapa de solos gerado pela técnica de árvore de decisão.

MATERIAL E MÉTODOS

Área de estudo

A área de estudo está definida pelo polígono do município de Erechim, no Planalto Médio do Rio Grande do Sul, entre as coordenadas SIRGAS2000 52°06’ e 52°24’ de longitude Oeste e 27°28’ e 27°47’ de latitude Sul. A área compreende 889,33 km² e possui uma elevação local que varia entre 423 a 900 m (Figura 1). A região foi escolhida por possuir mapas de solos apenas em nível de reconhecimento e exploratório (ten Caten et al., 2007), fazendo dessa uma região carente por informação espacial em solos.

Visando gerar as AD, as relações solo-paisagem foram obtidas a partir de reambulação da área de estudo com informações obtidas no campo por um pedólogo experiente. Foram descritos perfis e coletadas amostras de solos para análise em laboratório. Essas informações possibilitaram delinear as unidades de mapeamento em cartas topográficas de escala 1:50.000 e definir as classes de solos de ocorrência significativa de acordo com o Sistema Brasileiro de Classificação de Solos - SiBCS (EMBRAPA, 2006) (Tabela 1). Essas informações, em conjunto com as covariáveis ambientais, foram utilizadas como dados de treinamento para as AD.

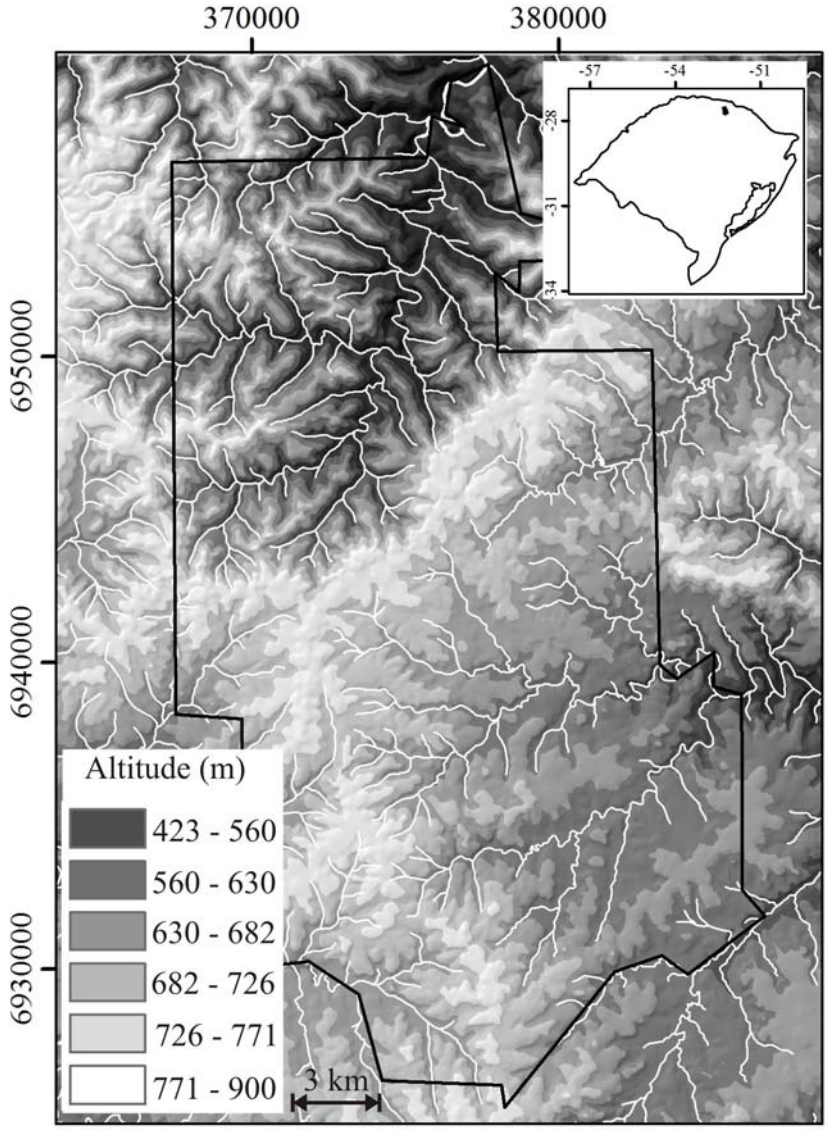


Figura 1: Área do estudo com o perímetro do município de Erechim sobreposto à rede de drenagem e ao mapa hipsométrico da região. Imagem interna, no canto superior direito, localiza a área do estudo no Estado do Rio Grande do Sul.

Covariáveis ambientais

Neste estudo, foram utilizadas covariáveis do modelo 'scorpan' (McBratney et al, 2003), relacionadas aos fatores de formação do solo relevo (*r*) e organismos (*o*). As demais covariáveis do modelo não estão disponíveis ou não possuem a resolução espacial demandada para este estudo. O fator organismos foi representado pela covariável Índice de Vegetação por Diferença Normalizada (NDVI) do mês de outubro (NOUT) e pelo desvio padrão do NDVI

(DEPA). A opção pelo NDVI do mês de outubro foi devido a este ser o mês de maior contraste entre os diferentes usos da terra na região. Para o cálculo de ambas as covariáveis, as imagens utilizadas são todas do ano de 2004 devido à maior disponibilidade de imagens com ausência de nuvens durante aquele ano. Todos os cálculos de NDVI foram realizados conforme Jensen (2009).

Tabela 1 – Classes de solos encontradas na área de estudo.

Classe SiBCS - legenda	Área (ha)	Fração total (%)
GLEISSOLO HÁPLICO - GX	31,28	12,52
LATOSSOLO VERMELHO - LV	122,99	49,22
NEOSSOLO LITÓLICO - RL	26,26	10,51
Associação NEOSSOLO LITÓLICO e CAMBISSOLO HÁPLICO – RL/CX	18,27	7,31
Associação NITOSSOLO VERMELHO e CAMBISSOLO HÁPLICO – NV/CX	7,05	2,82
Associação NITOSSOLO VERMELHO e CHERNOSSOLO ARGILÚVICO fase altitude elevada – NV/MT fae	26,01	10,41
Associação NITOSSOLO VERMELHO e CHERNOSSOLO ARGILÚVICO fase altitude baixa – NV/MT fab	18,03	7,22

O preditor DEPA foi gerado a partir de dados de oito distintas datas do ano de 2004, obtidas pela plataforma Landsat 5 sensor TM com resolução espacial de 30 m. Cada data teve seu valor de NDVI calculado individualmente. Em seguida, foi executado o cálculo do desvio padrão do valor do NDVI entre as oito datas para cada pixel no programa ArcGIS (ESRI, 2008), na função Raster Calculator. O desvio padrão foi calculado como segue:

$$DEPA = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(NDVI_i - \overline{NDVI}\right)^2}$$

onde

$DEPA$ = desvio padrão do NDVI;

n = número de distintas datas utilizadas para gerar o DEPA;

$NDVI_i$ = iéssimo NDVI;

$\overline{NDVI}$ = valor médio do NDVI.

O fator relevo foi representado pelas covariáveis elevação (ELEV), distância vertical da rede de drenagem (DIST), declividade (DECL), índice de umidade topográfica (IUT), capacidade de transporte de sedimento (CTS), curvatura planar (PLAN), curvatura de perfil (PERF) e rugosidade (RUGO). As covariáveis foram geradas de acordo com Wilson e Gallant (2000), a partir de um modelo digital de elevação (MDE). O MDE com resolução espacial de 30 m foi derivado de curvas de nível de cartas topográficas com escala 1:50.000. A interpolação das curvas de nível para o formato matriz ocorreu na ferramenta *topo do raster* do programa ArcGIS 9.3 (ESRI, 2008), usando a técnica *spline* (Wahba, 1990). No programa SAGA-GIS (Olaya, 2004), os atributos ELEV, DIST, DECL, PLAN, PERF, e RUGO foram gerados na ferramenta *standard terrain analysis*, e os atributos IUT e CTS com a ferramenta *grid calculator*.

A tabulação dos dados para tratamento estatístico ocorreu a partir da amostragem do mapa de solos de treinamento sobreposto aos dez planos de informação das covariáveis ambientais. Foi gerada uma tabela com 277.491 linhas e 11 colunas, na qual estão representados 249,80 km² de superfície, ou seja, 28% da área total do município de Erechim.

A tabela foi submetida à análise estatística multivariada para a verificação de presença de multicolinearidade entre os preditores, tendo sido executado o teste de correlação de

Pearson e Análise de Componentes Principais (ACP) (Hair et al., 2006). A análise de correlação e a ACP foram executadas no programa R (R DEVELOPMENT CORE TEAM, 2011).

Árvores de decisão

A tabela com 28% dos dados originais foi reamostrada em sete conjuntos de dados. Os conjuntos corresponderam às densidades de amostragem de um, três, cinco, 10, 15, 20 e 25% da área total do estudo. Todas as classes de solos foram amostradas proporcionalmente a sua área, conforme Scull et al. (2005). O desenvolvimento das AD foi realizado no programa de mineração de dados WEKA (Hall et al., 2009). Para o processamento dos dados, foi utilizado o algoritmo J48 devido a esse ter tido o melhor desempenho em estudo realizado por Giasson et al. (2011). O número mínimo de observações (*minNumObj* no WEKA) por folha para cada conjunto de dados foi determinado após uma análise do percentual de observações erroneamente classificadas. Essa análise foi executada com uma série de valores para o número mínimo de observações em cada um dos sete conjuntos de dados. O método de poda, que mitigava o erro na árvore gerada, também foi selecionado (*reducedErrorPruning = True* no WEKA). Durante a fase de geração da árvore, foi utilizado um conjunto de dados independente de cinco por cento da área total (*Supplied test set* no WEKA), com o intuito de verificar a qualidade da árvore gerada.

Mapa de solos

As AD que apresentaram a melhor relação custo-benefício entre a complexidade do modelo e o número de observações erroneamente classificadas foram implementadas no programa ArcGIS, na função *raster calculator*. As informações derivadas da árvore foram convertidas na função condicional ‘con(teste, verdadeiro, falso)’ do programa. Essa função permite que os arquivos raster com as covariáveis ambientais sejam processados de acordo com o conjunto de regras derivadas da AD. Como a escala de publicação pretendida é de

1:50.000, em cada mapa de solos gerado, foram extraídas regiões de píxeis isolados com área mínima mapeável menor do que um hectare.

Qualidade dos modelos e mapas

A qualidade dos modelos AD foi avaliada a partir do valor percentual de observações erroneamente classificadas em toda a árvore. Esse valor é uma das saídas do programa WEKA após o conjunto de dados teste ser processado no modelo gerado. Para seu cálculo, o programa soma todas as observações erroneamente classificadas e as divide pelo total de observações do conjunto teste, por fim multiplica por 100 para gerar o valor percentual (Hall et al., 2009).

A precisão do modelo ajustado foi avaliada pelo conjunto de classes corretamente classificadas e divididas pelo somatório do conjunto de classes correta e erroneamente classificadas. Em seguida, esse valor é multiplicado por 100 para se ter um valor relativo (Hall et al., 2009).

O índice kappa foi utilizado para verificar a qualidade da informação contida no mapa (Hengl et al., 2007). A matriz dos erros para o cálculo do índice kappa foi executada conforme Congalton (1991), a partir de reambulação a campo para identificação de 50 locais, por classe, do solo ocorrente. No total, foram identificados 350 locais de ocorrência das classes de solos mapeadas neste estudo.

RESULTADOS E DISCUSSÃO

Estudo da multicolinearidade entre os preditores

A análise de correlação nas 277.491 linhas de dados pode ser visualizada na Tabela 2. Apenas a correlação entre PERF e DECL mostrou-se não significativa. As covariáveis DEPA e NOUT mostraram ter correlação negativa (-0,480). A existência dessa correlação está ligada ao fato de que o NDVI do mês de outubro (NOUT) foi utilizado para o cálculo do desvio padrão do ano de 2004 (DEPA). Além disso, no mês de outubro, ocorrem grandes contrastes

dos valores de NDVI na paisagem. As áreas com mata irão apresentar elevados índices de vegetação, ao passo que as áreas cultivadas com a cultura do trigo senescida ou colhida irão apresentar os menores valores para o índice de vegetação. Estes locais de grande variação do NDVI ao longo do ano geram os maiores valores da covariável DEPA. Os menores valores da covariável NOUT encontram-se espacialmente nos mesmos locais onde DEPA apresenta seus maiores valores, por isso a correlação negativa.

Tabela 2 – Matriz de correlação entre covariáveis ambientais.

	DEPA	NOUT	CTS	IUT	ELEV	PLAN	PERF	DECL	DIST	RUGO
DEPA	1									
NOUT	-0,480[a]	1								
CTS	-0,197[a]	0,172[a]	1							
IUT	0,049[a]	-0,066[a]	0,122[a]	1						
ELEV	0,129[a]	-0,113[a]	-0,209[a]	-0,152[a]	1					
PLAN	0,057[a]	-0,058[a]	-0,345[a]	-0,336[a]	0,060[a]	1				
PERF	0,076[a]	-0,044[a]	-0,131[a]	-0,170[a]	0,290[a]	0,286[a]	1			
DECL	-0,335[a]	0,302[a]	0,345[a]	-0,515[a]	-0,175[a]	0,052[a]	0,005[b]	1		
DIST	-0,132[a]	0,142[a]	-0,014[a]	-0,416[a]	0,278[a]	0,244[a]	0,428[a]	0,368[a]	1	
RUGO	-0,291[a]	0,245[a]	0,278[a]	-0,294[a]	-0,150[a]	0,018[a]	-0,021[a]	0,715[a]	0,250[a]	1

[a]: $p < 0.001$; [b]: não significativo. NDVI do mês de outubro de 2004 (NOUT), desvio padrão do NDVI no ano de 2004 (DEPA), elevação (ELEV), distância vertical da rede de drenagem (DIST), declividade (DECL), índice de umidade topográfica (IUT), capacidade de transporte de sedimento (CTS), curvatura planar (PLAN), curvatura de perfil (PERF) e rugosidade (RUGO).

A maior correlação foi verificada entre as covariáveis RUGO e DECL (0,715). A existência de correlação entre essas variáveis se deve ao fato de que a declividade é uma das variáveis utilizadas para o cálculo da rugosidade, assim como os maiores valores de ambas as covariáveis foram verificados espacialmente nos mesmos locais, disto a correlação positiva.

O resultado da rotação dos eixos originais, visando potencializar a variabilidade em um novo conjunto de variáveis, pode ser visualizado na Tabela 3. Cerca de 78% da variância original fica distribuída entre as cinco primeiras componentes, com percentuais consideráveis de variabilidade retida entre a sexta e a nona componente. Valores residuais de variância nas componentes mais elevadas é um indicativo de um relativo grau de independência entre as variáveis originais.

Tabela 3 – Resultado da distribuição da variabilidade nas dez componentes principais.

Componente Principal	Autovalor	Variância (%)	Variância acumulada (%)
1	2,728	27,276	27,276
2	2,189	21,892	49,168
3	1,105	11,046	60,215
4	1,032	10,324	70,539
5	0,779	7,794	78,333
6	0,537	5,367	83,700
7	0,510	5,104	88,804
8	0,483	4,831	93,635
9	0,441	4,408	98,043
10	0,196	1,957	100,000

Resultados similares foram identificados por Hengl e Rossiter (2003), que utilizaram de ACP para mitigar o efeito de multicolinearidade entre atributos do terreno. Os autores identificaram 80% da variabilidade retida nas cinco primeiras componentes geradas a partir da rotação ortogonal de um conjunto de nove atributos do terreno. Para os autores, esse resultado demonstrou que as covariáveis eram menos redundantes do que se havia inicialmente pressuposto. Em estudo realizado por ten Caten et al. (2011), as três primeiras componentes principais, as quais reteram 65% da variabilidade original dos dados, possibilitaram a geração de modelos preditivos com desempenho similar àquelas que se utilizaram da totalidade das

covariáveis preditoras, com a vantagem de demandar que um menor número de dados fosse manipulado.

Uma visão mais detalhada das correlações entre as covariáveis pode ser obtida analisando-se a distribuição dessas no diagrama de ordenação unitário para a duas primeiras componentes principais (Figura 2). Nessas componentes, foi retida cerca de 50% da variabilidade original dos dados. A proximidade entre as covariáveis DECL e RUGO, bem como entre PLAN, PERF e ELEV indica haver entre elas uma redundância mais intensa do que em meio às covariáveis espacialmente mais distantes no gráfico. A posição do IUT reproduz o que pode ser verificado na Tabela 1, ou seja, a covariável é inversamente correlacionada com praticamente todas as covariáveis, à exceção da CTS e DEPA.

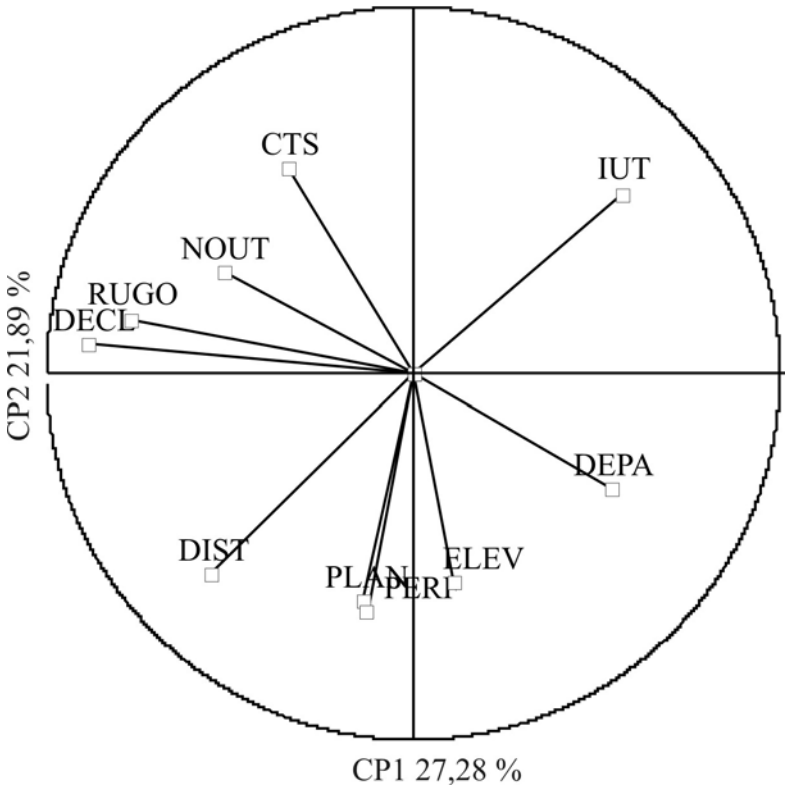


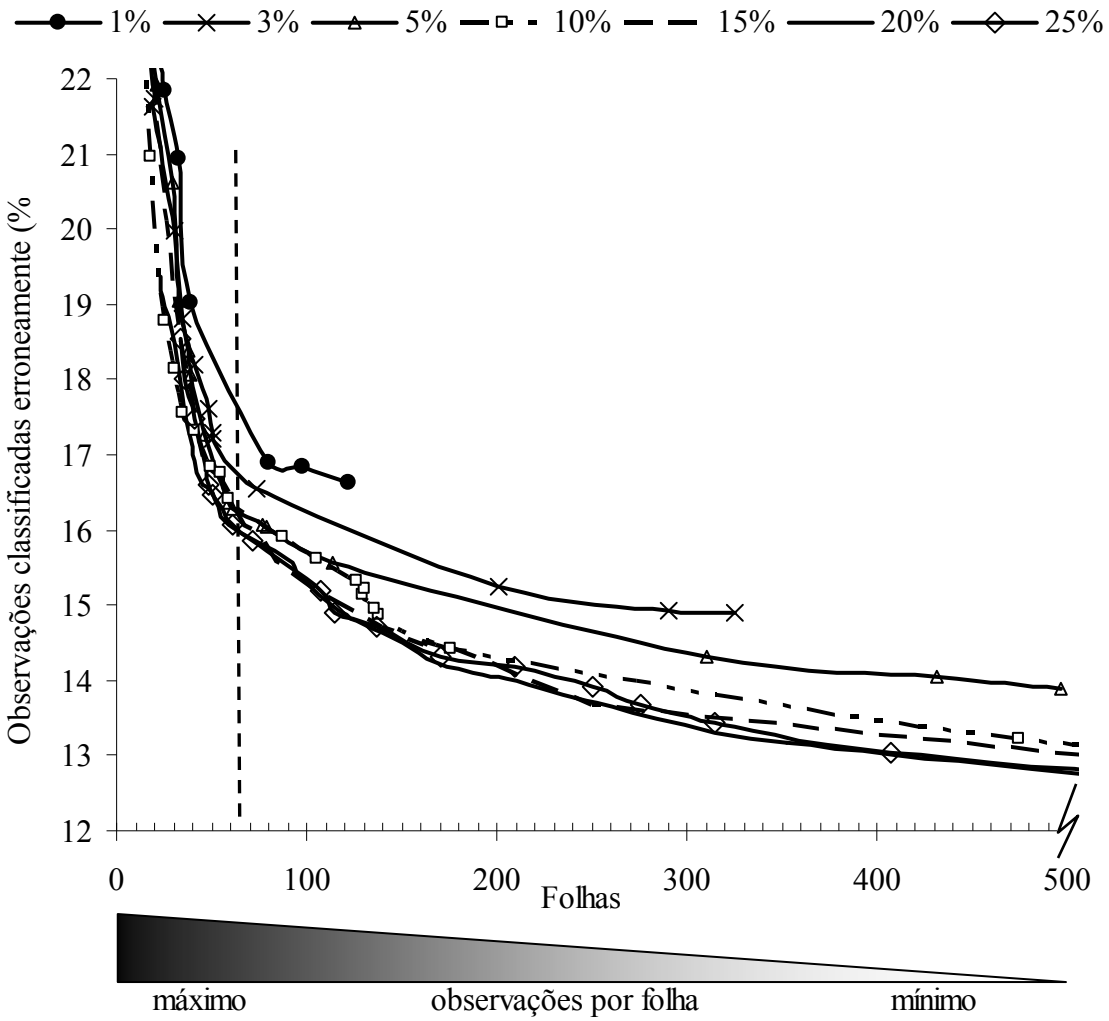
Figura 2 – Diagrama de ordenação unitário com as coordenadas das covariáveis ambientais para as duas primeiras componentes principais. NDVI do mês de outubro de 2004 (NOUT), desvio padrão do NDVI no ano de 2004 (DEPA), elevação (ELEV), distância vertical da rede

de drenagem (DIST), declividade (DECL), índice de umidade topográfica (IUT), capacidade de transporte de sedimento (CTS), curvatura planar (PLAN), curvatura de perfil (PERF) e rugosidade (RUGO).

A partir do estudo do padrão da variabilidade presente nas covariáveis preditoras, decidiu-se simplificar o conjunto de dados utilizado para gerar os modelos por árvore de decisão. Foram descartadas da sequência do estudo as covariáveis PLAN, PERF, RUGO e NOUT.

Complexidade do modelo por árvore de decisão

AD com um grande número de folhas se superajustam aos dados e atingem seu máximo potencial preditivo (Figura 3). Disso decorre que, durante o desenvolvimento dos modelos por árvore de decisão, faz-se necessário avaliar o ganho preditivo dos modelos na medida em que esses se tornam mais complexos. Para os sete conjuntos de amostras, verifica-se que na medida em que as AD tornam-se mais complexas, com um maior número de folhas, menor é o número de observações erroneamente classificadas pelos modelos. No entanto, mesmo para árvores muito complexas e maiores volumes de dados para ajustar os modelos, o percentual de observações erroneamente classificadas permanece em torno de 13% (Figura 3). Possivelmente esse fato decorra da impossibilidade de se explicar a complexidade da distribuição espacial das classes de solos na região apenas com as seis covariáveis utilizadas. Scull et al. (2005) observaram que em torno de 18,52% das observações localizadas na área de treinamento e 24,34% daquelas localizadas na área de validação eram classificadas erroneamente na fase de geração de modelos de AD para o nível de ordem.



240

241 Figura 3 – Relação entre o número de folhas na árvore de decisão e o percentual de amostras
242 erroneamente classificadas.

243 Para os conjuntos de dados de um e três por cento, mesmo com o menor número de
244 observações por folha em árvores mais complexas, não são alcançados resultados melhores do
245 que 16,5 e 15%, respectivamente. Entre os efeitos de se utilizar um número pequeno de
246 amostras para a geração das AD, pode estar a não representatividade desses em relação ao
247 conjunto total de dados disponíveis, acarretando em um menor poder preditivo.

248 À medida que as AD são podadas, ocorre um aumento no número de observações
249 erroneamente classificadas (Figura 3). No entanto, a partir de um determinado número de
250 folhas, a taxa de incremento dos erros modifica-se sensivelmente e passa a aumentar

linearmente para AD com um número de folhas menor do que 65 (linha vertical tracejada). Esse comportamento também é verificado nos menores conjuntos de dados de um e três por cento, embora aconteça de maneira não tão marcante como nos demais conjuntos de dados.

Buscando utilizar modelos que associassem simplicidade e poder preditivo nos diferentes conjuntos de dados, optou-se por utilizar o valor de 65 folhas como limite mínimo para o número de folhas nas AD. A partir desse valor, foram geradas AD para os conjuntos de dados de cinco, 10, 15, 20 e 25% (Tabela 4). Os conjuntos de dados de um e três por cento foram descartados por apresentarem os piores resultados quanto ao número de observações erroneamente classificadas na região de AD, com 65 folhas. Grinand et al. (2008) afirmam que volumes de amostras menores do que 10% podem levar a uma redução substancial da qualidade dos modelos.

Tabela 4 – Informações de saída do programa WEKA a partir do teste dos cinco modelos no conjunto de dados independentes (*Supplied test set*).

Conjuntos de dados (%)	Observações classificadas erroneamente (%)	Kappa (%)	Precisão (%)
5	16,27	76,50	83,00
10	16,23	76,80	83,20
15	16,14	76,94	83,30
20	15,79	77,36	83,70
25	16,02	77,06	83,60

Os resultados da Tabela 4 indicam que houve uma tendência para uma pequena diminuição do número de observações erroneamente classificadas à medida que o conjunto de dados utilizado para gerar os modelos era aumentado. O fato de o número de observações incorretas permanecer em torno de 16% para todos os cinco conjuntos de dados, possivelmente se deva à qualidade e quantidade das covariáveis preditoras utilizadas neste estudo, as quais foram incapazes de predizer a total complexidade da distribuição espacial da classes de solos na área de estudo. Os valores do índice Kappa e de precisão dos modelos gerados pelo programa WEKA também apresentaram sensível melhora à medida que um

maior volume de dados foi utilizado para gerar os modelos (Tabela 4). Contudo, esses resultados indicam que a melhoria do ajuste dos modelos gerados por AD deverá vir associada a outros fatores, tais como a utilização de uma maior quantidade e diversidade de covariáveis preditoras, e não somente a utilização de um maior volume de dados para ajustar os modelos.

Em estudo realizado por Giasson et al. (2011), os autores identificaram um índice kappa de 57,1% para uma AD com 79 folhas, sendo utilizada para mapear seis classes de solos. De acordo com esses autores, AD muito simplificadas implicavam um reduzido poder preditivo dos modelos. Naquele estudo, foi avaliada uma AD com apenas cinco folhas, implicando um índice kappa de aproximadamente 43,9% e com apenas duas classes de solos sendo preditas.

Mapas de solos

As cinco AD foram utilizadas para gerar mapas de solos para o município de Erechim (Figura 4). Todos os cinco mapas de solos reproduziram a relação solo-paisagem, formalizada pelo pedólogo de maneira muito similar. Nas colinas localizadas em locais de maior altitude, foram posicionados os Latossolos Vermelhos (Figura 4). Nesses locais, as condições de relevo possibilitaram a formação desses solos profundos, homogêneos e bem intemperizados.

Em posições mais baixas na toposequência, foram espacializadas as associações entre Neossolos Litólicos e Cambissolos Háplicos (Figura 4). Essa associação encontra-se ao longo de toda a parte superior do vale do Rio Dourado. Nesses locais, a declividade possibilita a remoção constante de material. Como consequência, tem-se a presença de solos rasos e com grande presença de pedregosidade. Os Neossolos Litólicos foram espacializados, predominantemente, nas rampas mais longas e declivosas da toposequência (Figura 4). A associação Nitossolo Vermelho e Chernossolo Argilúvico fase altitude baixa foi espacializada ao longo da drenagem dos rios nas posições encaixadas dos vales. Nesses locais, encontram-

se áreas planas que possibilitaram a formação desses solos mais profundos e desenvolvidos (Figura 4).

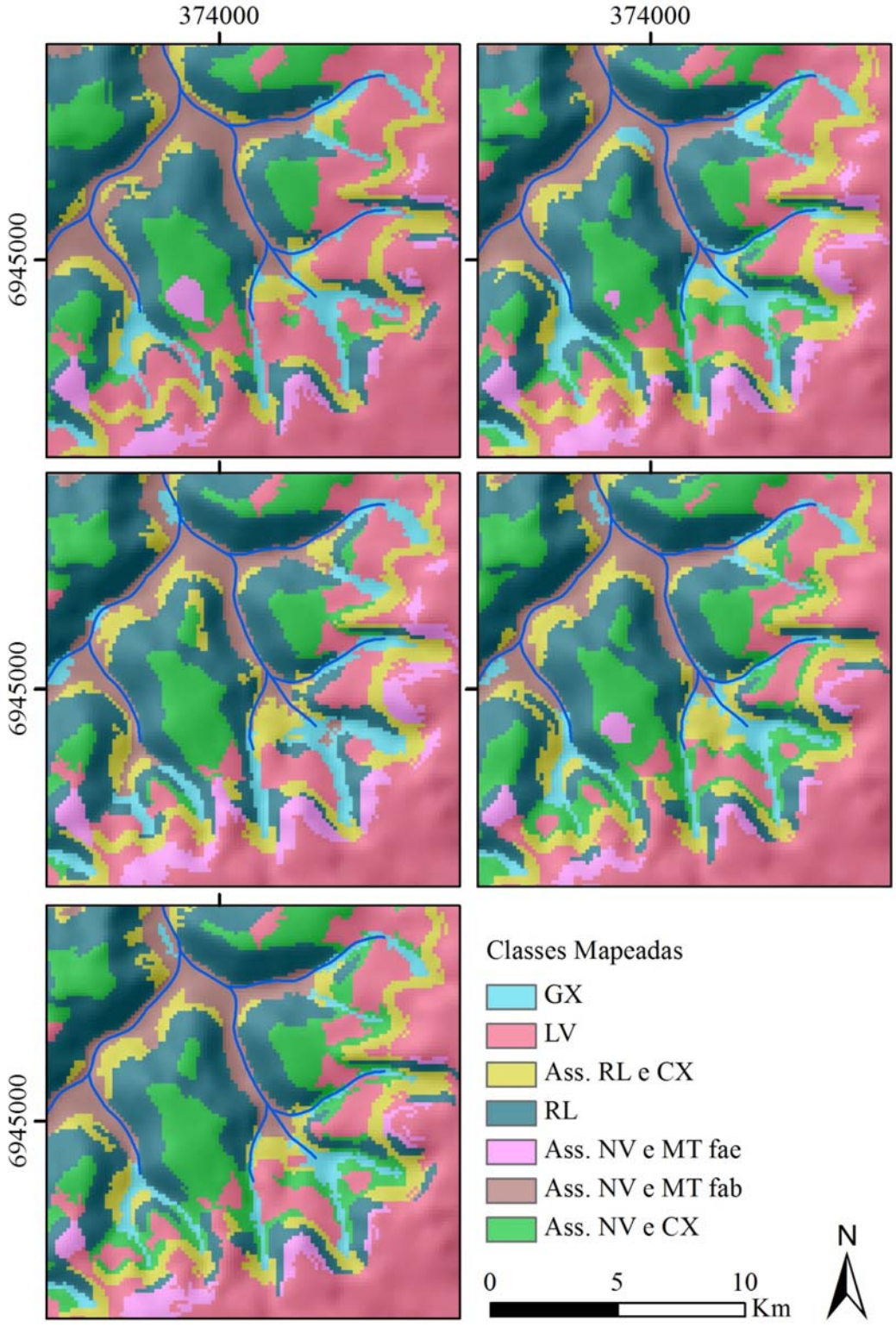


Figura 4 – Sessões dos mapas de solos preditos a partir dos cinco conjuntos de amostras a) 5%, b) 10%, c) 15%, d) 20%, e) 25%. Coordenadas UTM Fuso 22 / SIRGAS2000.

A análise visual dos cinco mapas gerados não fornece subsídios suficientes para a decisão de qual densidade de dados é a mais adequada. Na sequência do estudo, fez-se a avaliação em campo da capacidade preditiva de cada uma das AD geradas a partir dos cinco conjuntos de dados.

Qualidade dos mapas de solos

A partir dos cinco mapas de solos gerados e dos dados coletados a campo, foi possível avaliar a qualidade dos modelos preditivos (Tabela 5). Optou-se, primeiramente, por avaliar a semelhança entre os mapas gerados a partir da comparação entre eles. Verifica-se que, à medida que o volume de dados utilizado aumenta, maior também é a concordância entre os mapas gerados. Quando mapas com volumes de dados de 5 e 10% são comparados, o índice kappa é menor do que o índice obtido da comparação entre mapas com volumes de dados de 20 e 25%. Esse resultado é um indicativo de que volumes de dados maiores possibilitam capturar melhor a complexidade a ser mapeada, embora as semelhanças encontradas entre os mapas preditos com volumes menores de dados, neste estudo, foram também elevadas.

Tabela 5 – Índice kappa entre os mapas gerados com diferentes densidades de amostragem, com a área de treinamento e com amostras no campo.

	5%	10%	15%	20%	25%
5%	-	-	-	-	-
10%	79,4	-	-	-	-
15%	78,4	81,1	-	-	-
20%	81,1	84,1	84,1	-	-
25%	82,4	85,3	85,6	90,4	-
treinamento	77,3	77,3	77,1	77,7	77,6
campo	69,7	68,5	71,1	71,4	71,1

A reprodutibilidade dos cinco modelos por AD empregados neste estudo pode ser visualizada na sexta linha da Tabela 5. Observa-se que todos os modelos reproduziram, com

um valor superior a 77%, a classe de solo que havia sido definida pelo pedólogo nas áreas de treinamento dos modelos. Esse valor é superior aos 46,12% encontrados por ten Caten et al. (2011), que utilizaram de regressões logísticas múltiplas como modelos preditivos.

Embora o presente estudo tenha uma reprodutibilidade elevada, em aproximadamente 23% da área de treinamento, os modelos divergiram do que havia sido definido como sendo a classe de solo ocorrente. Esse fato pode estar ligado a erros na descrição do solo a campo ou na incapacidade dos modelos em capturar a complexidade inerente à distribuição espacial das classes de solos no local.

A acurácia dos modelos preditivos está representada na última linha da Tabela 5. Todos os modelos possibilitaram uma acurácia próxima a 70% no mapeamento das classes de solos locais. Como era esperado, a acurácia dos modelos preditivos é um pouco inferior à reprodutibilidade dos modelos comentada nos parágrafos anteriores. Contudo, ambos os indicadores da qualidade do mapeamento indicam que os cinco modelos testados desempenham de maneira satisfatória o mapeamento de classes de solos.

Quanto à influência do volume de dados utilizados para gerar as AD, o que se observa é que há uma tendência de melhor desempenho dos modelos preditivos quando o volume de dados utilizado para treinar os modelos é maior (Tabela 5). O melhor desempenho preditivo foi atingido pelos volumes de dados de 15, 20 e 25%. Estudo realizado por Grinand et al. (2008) testou amostras de 10 a 90% dos dados originais. Os autores verificaram que a grande variação da acurácia estava entre 10 e 20% dos dados, não sendo verificado significativo aumento de qualidade dos modelos a partir de 30% do volume total de amostras. De acordo com os resultados aqui encontrados, volumes de dados de 15% aliariam um menor volume de dados, reprodutibilidade e acurácia no mapeamento.

O mapeamento de classes de solos na carta topográfica de Dois Córregos (SP) foi executado por Crivelenti et al. (2009) com um volume de dados de 714.000 amostras em uma

área total do estudo de 772 km², e embora não tenha sido possível a determinação exata do volume de dados utilizados para a geração das árvores de decisão naquele estudo, acredita-se que mais de 80% dos dados originais tenham sido utilizados para gerar os modelos. Os autores reportaram um índice kappa de 43% em relação ao mapa utilizado para gerar o modelo.

Estudos realizados por Bui e Moran (2003) demonstraram que a capacidade preditiva dos modelos experimentou um significativo incremento a partir da utilização de 10% dos dados originais. A qualidade dos modelos preditivos teve um valor máximo quando as amostras utilizadas representaram 25% do total de dados, com um índice kappa de 64% em relação ao mapa original. Contudo, os autores afirmam que a menor densidade de amostragem, de 10%, já seria suficiente para capturar grande parte da diversidade contida nos dados para a construção dos modelos preditivos por AD.

CONCLUSÃO

O presente estudo demonstrou que conjuntos de dados pouco representativos (menores do que 5%), para a geração de modelos de Árvore de Decisão, têm dificuldades em capturar a complexidade existente na distribuição espacial de solos. Por outro lado, conjuntos de dados superiores a 20% podem implicar em redundância excessiva para a geração de modelos preditivos, implicando manipulação desnecessária de dados, o que não irá refletir em ganho preditivo pelos modelos derivados desses conjuntos de dados. Nas condições em que o presente estudo foi realizado, volumes de dados entre cinco e 15% implicaram na melhor relação custo-benefício quanto ao volume de dados a ser manipulado e à capacidade preditiva dos modelos gerados. Valores de reprodutibilidade em torno de 77% e acurácia próxima a 70% foram alcançados com esses volumes de amostras utilizados para gerar os modelos por árvore de decisão.

AGRADECIMENTOS

À CAPES pelos recursos utilizados na execução do trabalho e ao CNPq pela bolsa de produtividade em pesquisa do segundo autor e aporte financeiro para auxílio desse trabalho.

REFERÊNCIAS

Bui, E. N.; Moran, C. J. 2003. A strategy to fill gaps in soil survey over large spatial extents: an example from the Murray-Darling basin of Australia. Geoderma 111: 21-44.

Crivelenti, R.C. et al. 2009. Mineração de dados para a inferência de relações solo-paisagem em mapeamentos digitais de solo. Revista Agropecuária Brasileria. 44: 1707-1715.

Congalton, R.G. 1991. A review of assessing the accuracy of classification of remotely sensed data. Remote Sensing of Environment 37: 35-46.

Empresa Brasileira de Pesquisa Agropecuária [EMBRAPA]. 2006. Sistema brasileiro de classificação de solos. Centro Nacional de Pesquisa em Solos, Rio de Janeiro, RJ, Brasil.

Environmental Systems Research Institute [ESRI], Inc., Redlands, CA, 2008.

Giasson, E.; Sarmento, E.C.; Weber, E.; Flores, C.A.; Hasenack, H. 2011. Decision trees for digital soil mapping on subtropical basaltic steeplands. Scientia Agrícola 68: 167-174.

Geissen, V.; Kampichler, C.; Lopez-de Llergo-Juarez, J.J.; Galindo-Acantara, A. 2007. Superficial and subterranean soil erosion in Tabasco, tropical Mexico: Development of a decision tree modeling approach. Geoderma 139: 277-287.

Grinand, C.; Arrouays, D.; Laroche, B.; Martin, M. 2008. Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context. Geoderma 143: 180-190

Grunwald, S. 2009. Multi-criteria characterization of recent digital soil mapping and modeling approaches. Geoderma 152: 195-207.

Hair, J.F.; Anderson, R.E.; Tatham, R.L.; Black, W. C. 2006. Análise Multivariada de Dados. 5ed. Bookman, Porto Alegre, RS, Brasil.

399 Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. 2009. The
400 WEKA Data Mining Software: An Update; SIGKDD Explorations, 11.
401 Hengl, T.; Rossiter, D.G. 2003. Supervised Landform classification to enhance and replace
402 photo-interpretation in semi-detailed soil survey. Soil Science Society of America Journal 67:
403 1810-1822.
404 Hengl, T.; Toomanian, N.; Reuter, H.I.; Malakouti, M.J. 2007. Methods to interpolate soil
405 categorical variables from profile observations: Lessons from Iran. Geoderma 140: 417-427.
406 Jensen, J.R. 2009. Sensoriamento Remoto do Ambiente: uma perspectiva em recursos
407 terrestres. Ed. Parêntese, São José dos Campos, SP, Brasil.
408 Kheir, R.B. et al. 2010a. Predictive mapping of soil organic carbon in wet cultivated lands
409 using classification-tree based models: The case study of Denmark. Journal of Environmental
410 Management 91: 1150-1160.
411 Kheir, R.B. et al. 2010b. Spatial soil zinc content distribution from terrain parameters: A GIS-
412 based decision-tree model in Lebanon. Environmental Pollution 158: 520-528.
413 Lemercier, B.; Lacoste, M.; Loum, M; Walter, C. 2011. Extrapolation at regional scale of
414 local soil knowledge using boosted classification trees: A two-step approach, Geoderma, In
415 Press, Corrected Proof.
416 McBratney, A. B.; Mendonça-Santos, M. L.; Minasny, B. 2003. On digital soil mapping.
417 Geoderma 117: 3-52.
418 Olaya, V. 2004 A gentle introduction to SAGA GIS. The SAGA user group, Göttingen,
419 Niedersachsen, Germany.
420 Qi, F.; Zhu, A. X. 2003. Knowledge discovery from soil maps using inductive learning.
421 International Journal of Geographical Information Science 17: 771-795.
422 R Development Core Team. 2011. R: A Language and Environment for Statistical
423 Computing, R Foundation for Statistical Computing, Vienna, Vienna, Austria.

424 Scull, P.; Franklin, J.; Chadwick, O.A. 2005. The application of classification tree analysis to
425 soil type prediction in a desert landscape. Ecological Modelling 181: 1-15.
426 ten Caten, A; Dalmolin, R.S.D.; Pedron, F.A.; Mendonça-Santos, M. de L. 2011. Estatística
427 multivariada aplicada à diminuição do número de preditores no mapeamento digital do solo.
428 Pesquisa Agropecuária Brasileira 46: 554-562.
429 ten Caten, A. ; Silva, C.A. ; Dalmolin, R. S. D. 2007. Mapas disponíveis e a demanda por
430 levantamentos de solos em grandes escalas. In: Congresso Brasileiro de Ciência do Solo. 31.
431 Anais. Gramado. CD-ROM.
432 Wilson, J. P. e Gallant, J. C. 2000. Digital terrain analysis. p. 1-27. *In:* Wilson, J. P. e Gallant,
433 J. C. eds. Terrain analysis: principles and applications. Wiley & Sons, New York, NY, USA.
434 Witten, I.H. e Frank, E. 2005. Data Mining: Practical Machine Learning Tools and
435 Techniques. 2ed. Morgan Kaufmann, San Francisco, Califórnia, USA.
436 Wahba, G. 1990. Spline models for Observational data. Paper presented at CBMS-NSF
437 Regional Conference Series in Applied Mathematics. Society for Industrial and Applied
438 Mathematics, Philadelphia, PA, USA.

DISCUSSÃO

Como esta tese foi estruturada em capítulos na forma de artigos, os quais constituem-se em si próprios em trabalhos completos, buscou-se nesta discussão traçar comentários gerais a respeito dos artigos, assim como a respeito do estado da arte do MDS no Brasil.

Em virtude do curto período de atividades, a pesquisa em MDS brasileira ainda não é resultado de grupos consolidados nessa linha de trabalho (Tabela 01 – Artigo 01). Disso decorre que, até o presente momento, pesquisadores têm se limitado a empregar estratégias automatizadas de mapeamento de solos como forma de se convencer e de convencer seus pares cientistas de que esta abordagem é capaz de gerar resultados similares àqueles alcançados pela metodologia convencional de mapeamento. A maioria das iniciativas de mapeamento de classes de solos tem se preocupado em avaliar a capacidade da abordagem automatizada em reproduzir o mapa que foi produzido pelo pedólogo, utilizando a abordagem convencional por fotointerpretação.

Como exemplo de mapeamento digital de classes de solos a partir de uma demanda, tem-se o estudo realizado por Silva (2011). O estudo detalha o mapeamento de classes de solos para o município de Erechim, a partir da necessidade de se espacializar as áreas aptas para o cultivo da erva mate. O autor utilizou de um mapa de classes de solos gerado a partir de modelos por Árvores de Decisão para gerar um mapa de aptidão agrícola das terras e de áreas propícias ao cultivo da erva mate. Segundo esse autor, a adequação dos usos da terra para o município de Erechim indicou a aptidão para atividades com lavouras, silvicultura e pecuária leiteira, com reduzido percentual de restrição de uso, que é reservado à preservação de fauna e flora.

Com os primeiros estudos publicados no país, a fase de 'nova abordagem' já pode ser tida como superada. Na pesquisa em MDS a ser realizada no país e especificamente para o mapeamento automatizado de classes de solos, três importantes frentes de estudos podem ser exploradas, a saber: (i) a identificação e o dimensionamento das demandas, otimizando os recursos disponíveis para gerar a informação espacial em solos; (ii) pontos metodológicos que influenciam a execução de mapeamentos de classes e propriedades de solos; (iii) a disponibilização das informações com a qualidade requerida e na forma que possa ser utilizada pelos usuários.

Analisando a identificação e dimensionamento de demandas, verifica-se que o MDS possui um problema em comum à abordagem convencional: a desinformação dos gestores e

tomadores de decisão. Esses desconhecem a aplicabilidade do mapa de solos para a melhoria da alocação e uso dos recursos naturais, econômicos e humanos. Por outro lado, a abordagem automatizada possui como vantagem o maior apelo visual e o potencial ilustrativo dos Sistemas de Informação Geográfica (Burrough et al., 1994). Assim, poderão auxiliar na divulgação e na demonstração de que mapas de classes de solos são ferramentas que qualificam a gestão e o planejamento (Bouma, 2005).

Se de um lado existe um grupo de usuários que até ignora a existência de mapas de solos, de outro existem usuários extremamente qualificados e ávidos pelo conhecimento da informação espacial em solos. Para pesquisadores de diversas áreas de conhecimento, a informação sobre o solo é fundamental para a execução de suas atividades científicas. Estudos relacionados à produção de sedimentos (Minella e Merten, 2011), impactos das mudanças do uso da terra no sequestro de carbono (Tornquist et al., 2009) e tomada de decisão em cenários de pouca disponibilidade de dados (Bacic et al., 2008) são alguns exemplos de linhas de pesquisa ávidas pelo conhecimento da distribuição espacial de propriedades e classes de solos. No entanto, na falta dessa informação, o que tem sido visto é que os estudos generalizam as informações existentes, muitas vezes em escalas inadequadas (Dalmolin et al., 2004). Também ocorre que pesquisadores de áreas não diretamente ligadas à ciência do solo buscam viabilizar as informações necessárias por conta própria, o que pode implicar em erros substanciais nas conclusões desses estudos. O pedólogo precisa viabilizar as informações necessárias à pesquisa na velocidade e qualidade demandadas, sob pena de ser dado como um profissional desnecessário em vários projetos.

Esta tese esteve focada na segunda frente de estudos sugerida ao MDS brasileiro, as quais são as questões metodológicas do mapeamento de classes de solos. Uma linha de trabalho muito importante para o MDS é o estudo da variabilidade espacial do solo, e nesse aspecto a escala e a resolução espacial refletem diretamente na capacidade de se compreender e estudar a variabilidade (Behrens et al., 2010). A análise Wavelet demonstra-se como uma abordagem promissora para o estudo e melhor entendimento do padrão de variabilidade de atributos de terreno. Trabalhos nessa linha devem explorar melhor não somente a relação entre as covariáveis preditoras e classes de solos, mas também como essa relação ocorre nas diferentes escalas. Como a contribuição de cada um dos fatores de formação do solo é distinta nas diversas escalas de estudo e representação da informação solo, assim também uma covariável preditora adequada para um estudo em uma escala 1:25.000 poderá não ser relevante para estudos em escala de 1:250.000. Sem a intenção de ser um trabalho conclusivo, o estudo realizado nesta tese buscou trazer e divulgar uma forma diferenciada de explorar a

variabilidade espacial. A análise Wavelet tem um enorme potencial para estudos de escalas espaço-temporais, especialmente onde um grande volume de dados esteja disponível e as relações sejam de complexo estabelecimento.

O grande potencial para coleta de dados ambientais faz necessário avaliar o melhor conjunto de amostras a ser utilizado para gerar o modelo preditivo (*instance selection*) (Schmidt et al., 2008). Na revisão de literatura, foi verificado que os estudos até aqui realizados no país não têm discriminado as amostras quanto ao seu potencial preditivo. Este estudo buscou, então, demonstrar que existem diferenças qualitativas entre as observações de acordo com a sua localização espacial relativa aos polígonos de solos definidos pelo conhecimento tácito do pedólogo. A estratégia de pré-processamento de dados apresentada é de simples implementação e possibilita que observações de covariáveis ambientais localizadas nas áreas de maior variabilidade espacial (bordas dos polígonos) sejam descartadas da fase de geração dos modelos.

Além da seleção das observações com maior potencial preditivo, a escolha entre as covariáveis que mais contribuem para a predição implica modelos mais eficientes e simplificados (*feature selection*) (Behrens et al., 2010). A Análise de Componentes Principais possibilita revelar as relações entre covariáveis preditoras e selecionar entre aquelas em que não ocorram efeitos de multicolinearidade (ten Caten et al., 2011a; ten Caten et al., 2011b).

A variabilidade da densidade de dados revelada pela revisão de literatura indicou que existe, ainda, falta de um número de referência para o volume de dados a ser manipulado na fase de ajuste dos modelos preditivos (Tabela 02 – Artigo 01). Este estudo indicou que valores entre cinco e 15% do universo total das observações (pixels) disponíveis seriam proporções de dados adequadas para o treinamento dos modelos preditivos. Sob a pena de, caso o valor seja menor, a complexidade das relações solo-paisagem não serem capturadas pelos modelos ou, caso a densidade de amostragem seja muito elevada, poder-se-á estar manipulando um excesso de dados. É importante ter em mente que esses percentuais dizem respeito a um universo total de 277.491 observações, totalizando 249,80 km², que estão distribuídos nos 889,33 km² daquele estudo. Esses valores de densidade de amostragem precisam ser mais bem avaliados nos casos em que a dimensão espacial do estudo não seja tão expressiva.

Ainda nessa linha de pesquisa, visando elucidar questões metodológicas para a implementação do MDS, existem muitas dúvidas a serem sanadas. Entre elas está a influência das diferentes fontes de Modelos Digitais de Elevação (MDE) na qualidade do fator relevo, e, consequentemente, nos atributos de terreno a serem gerados. O relevo tem sido um dos fatores

de formação mais utilizados pelo MDS (McBratney et al., 2003). A pesquisa precisa avaliar as novas tecnologias que irão possibilitar um ganho de resolução espacial nos atributos derivados do fator relevo. Isso ocorrerá especialmente em locais mais planos, onde o fator relevo não tem o mesmo poder preditivo do que nas áreas mais declivosas (Rivero et al., 2007).

A terceira frente de estudo, listada no segundo parágrafo dessa discussão, diz respeito à avaliação da qualidade da informação gerada. A abordagem mais comumente adotada tem sido a de comparar o mapa gerado com o mapa disponível pelas técnicas convencionais de levantamento. O índice kappa entre os dois mapas tem sido a forma mais frequente de indicar a qualidade do mapa gerado. A fidelidade da reprodução do mapa convencional pelo mapa gerado pelo MDS (índice kappa próximo de 100%) vem sendo utilizada como um indicativo de que a técnica de MDS é aceitável. Contudo, os erros inerentes ao mapa convencional não têm sido reportados, tão pouco a influência dos erros previamente existentes é estudada nos mapas gerados (Brus et al., 2011). Finalmente, a avaliação da qualidade do mapa gerado deverá ser testada a campo e reportada junto dele.

Quanto à forma de disponibilizar a informação, está claro que passa pelas tecnologias digitais (Grunwald, 2009). A produção de uma nova informação por vias automatizadas e sua posterior disponibilização somente pela impressão em meio analógico significa condenar essa informação ao mesmo destino de vários outros levantamentos de solos, hoje praticamente inacessíveis nas gavetas das instituições de ensino e pesquisa pelo país. Ao mesmo tempo em que é importante divulgar e tornar a nova informação acessível é importante que isso seja feito de maneira padronizada. Caso não sejam somados esforços em torno de uma normatização para a geração e a disponibilização dessa informação, corre-se o risco de, no futuro, existirem formatos de dados incompatíveis para a reunião em um mesmo projeto. O desafio de padronizar e uniformizar a forma de disponibilizar os dados não deve ser assumido por um único grupo de pesquisa, mas por vários pesquisadores, como os participantes da RedeMDS ou pela Divisão Especializada Solo no Espaço e no Tempo da Sociedade Brasileira de Ciência do Solo.

O presente estudo teve como foco o mapeamento de classes de solos, contudo o mapeamento de propriedades do solo deve ser foco maior de pesquisas. Devido ao caráter generalista dos mapas de classes de solos, pode estar no mapeamento de propriedades uma aproximação maior com o usuário final dessa informação. O mapeamento de propriedades como pH, textura, matéria orgânica e fertilidade pode possibilitar uma aplicação mais imediata da informação gerada, do que se comparado à informação presente em mapas de

classes de solos, as quais são por natureza generalizadas. A possibilidade inerente ao MDS de mapear a paisagem na forma de uma matriz regular de pontos (pixels) permite conhecer a informação solo de uma forma muito mais detalhada do que pode ser feito com os mapas de classes de solos (Zhu et al., 2001). Esse potencial do MDS precisa ser mais bem explorado no país, como também uma forma de divulgar mais a abordagem automatizada, tornando a informação sobre o solo mais clara e disponível ao usuário.

Decisiva para o futuro do MDS no país é a formação dos cientistas do solo, a qual necessariamente deverá abarcar uma nova gama de conhecimentos. Para Grunwald et al. (2011), além do conhecimento da teoria e prática da pedologia, o cientista do solo moderno deve (i) ser capaz de acessar e manipular as ferramentas geoespaciais, como o sensoriamento proximal e remoto do solo e de outras covariáveis ambientais, e (ii) ter fortes habilidades quantitativas, particularmente, em estatísticas e análise espacial.

REFERÊNCIAS BIBLIOGRÁFICAS

BACIC, I.L.Z.; ROSSITER, D.G.; MANNAERTS, C.M. Applicability of a distributed watershed pollution model in a data-poor environment in Santa Catarina State, Brazil. **Revista Brasileira de Ciência do Solo**, Viçosa, v. 32, p. 1699-1712, aug. 2008.

BEHRENS, T. et al. Multi-scale digital terrain analysis and feature selection for digital soil mapping. **Geoderma**, Amsterdam, v. 155, p. 175-185, mar. 2010.

BOUMA, J. Soil scientists in a changing world. **Advances in Agronomy**, Newark, v. 88, p. 67-96, nov. 2005.

BRUS, D.J.; KEMPEN, B.; HEUVELINK, G.B.M. Sampling for validation of digital soil maps. **European Journal of Soil Science**, v. 62, n. 3, p. 1365-2389, apr. 2011.

BURROUGH, P.A.; BOUMA, J.; YATES, S.R. The state of the art in pedometrics. **Geoderma**, Amsterdem, v. 62, p. 311-326, 1994.

DALMOLIN, R.S.D et al. Relação entre as características e o uso das informações de levantamentos de solos de diferentes escalas. **Ciência Rural**, Santa Maria, v. 34, n. 5, p. 1479-1486, set./out. 2004.

GRUNWALD, S. Multi-criteria characterization of recent digital soil mapping and modeling approaches. **Geoderma**, Amsterdam, v. 152, n. 3-4, p. 195-207, jul. 2009.

MCBRATNEY, A. B.; MENDONCA SANTOS, M. L.; MINASNY, B. On digital soil mapping. **Geoderma**, Amsterdam, v. 117, n. 1-2, p. 3-52, jun. 2003.

MINELLA, J.P.G.; MERTEN, G.H. Monitoramento de bacias hidrográficas para identificar fontes de sedimentos em suspensão. **Ciência Rural**, Santa Maria, v. 41, n. 3, p. 424-432, mar. 2011.

RIVERO, R.G.; S. GRUNWALD, S.; BRULAND, G.L. Incorporation of spectral data into multivariate geostatistical models to map soil phosphorus variability in a Florida wetland. **Geoderma**, Amsterdam, v. 140, p. 428-443, jun. 2007

SCHMIDT, K; BEHRENS, T.; SCHOLTEN, T. Instance selection and classification tree analysis for large spatial datasets in digital soil mapping. **Geoderma**, Amsterdam, v. 146, p. 138-146, jun. 2008.

SILVA, C. A. da **Zoneamento Pedoclimático da erva mate *Ilex paraguariensis* para o município de Erechim-RS.** 2011. 160f. Tese (Doutorado em Ciência do Solo)-Universidade Federal de Santa Maria, Santa Maria, 2011.

ten CATEN, A. et al. Componentes principais como preditores no mapeamento digital de classes de solos. **Ciência Rural**, Santa Maria, v. 41, n. 7, p. 1170-1176, jul. 2011a.

ten CATEN, A. et al. Estatística multivariada aplicada à diminuição do número de preditores no mapeamento digital do solo. **Pesquisa Agropecuária Brasileira**, Brasília, v. 46, n. 5, p. 554-562, maio 2011b.

TORNQUIST, C.G. et al. Spatially explicit simulations of soil C dynamics in Southern Brazil: Integrating century and GIS with i_Century. **Geoderma**, Amsterdam, v. 150, p. 404-414, apr. 2009.

ZHU, A.X. et al. Soil mapping using GIS, expert knowledge, and fuzzy logic. **Soil Science Society of America Journal**, Madison, v. 65, p. 1463-1472, set./oct. 2001.