UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

FACULDADE DE AGRONOMIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DO SOLO

TESTES METODOLÓGICOS PARA O MAPEAMENTO DIGITAL DE CLASSES DE SOLOS UTILIZANDO ÁRVORES DE DECISÃO

RODRIGO TESKE

(TESE)

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE AGRONOMIA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DO SOLO

TESTES METODOLÓGICOS PARA O MAPEAMENTO DIGITAL DE CLASSES DE SOLOS UTILIZANDO ÁRVORES DE DECISÃO

RODRIGO TESKE
Engenheiro Agrônomo (UDESC)

Tese apresentada como um dos requisitos à obtenção do Grau de Doutor em Ciência do Solo

Porto Alegre (RS) Brasil
Julho, 2014

CIP - Catalogação na Publicação

TESKE, RODRIGO
TESTES METODOLÓGICOS PARA O MAPEAMENTO DIGITAL DE CLASSES DE SOLOS UTILIZANDO ÁRVORES DE DECISÃO / RODRIGO TESKE. -- 2014.
86 f.

Orientador: ELVIO GIASSON.

Dissertação (Mestrado) -- Universidade Federal do Rio Grande do Sul, Faculdade de Agronomia, Programa de Pós-Graduação em Ciência do Solo, Porto Alegre, BR-RS, 2014.

1. CLASSES DE SOLO. 2. AVALIAÇÃO DE ACUÁRIA. 3. ATRIBUTOS DO TERRENO. 4. PERFIS DE SOLO. 5. CLASSIFICAÇÃO SUPERVISIONADA. I. GIASSON, ELVIO, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

RODRIGO TESKE

TESTES METODOLÓGICOS PARA O MAPEAMENTO DIGITAL DE CLASSES DE SOLOS UTILIZANDO ÁRVORES DE DECISÃO

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência do Solo da Faculdade de Agronomia da Universidade Federal do Rio Grande do Sul, como requisito parcial para a obtenção do título de Doutor em Ciência do Solo.

Aprovada em 10 de julho de 2014
Homologada em 14 de agosto de 2014.

BANCA EXAMINADORA:

Prof. Paulo César do Nascimento
UFRGS

Prof. Alexandre ten Caten
UFSC

Dr. Ivan Luiz Zilli Bacic
Epagri/Ciram

Orientador - Prof. Elvio Giasson
UFRGS

DEDICO este trabalho à minha mãe Ana Gloria Hillesheim pelo apoio e incentivo aos estudos.

AGRADECIMENTOS

À minha mãe Ana Gloria, pelos exemplos de dedicação, trabalho, ética, superação e apoio aos meus estudos e formação acadêmica.

Ao meu irmão Jackes e família, pelo apoio e incentivo.

À Universidade Federal do Rio Grande do Sul e ao Programa de Pós-Graduação em Ciência do Solo pela oportunidade de realização do curso de doutorado.

À CAPES pela concessão da bolsa de doutorado e ao CNPq à FAPERGS pelo financiamento dos projetos de pesquisa e apoio financeiro para participação e divulgação dos resultados deste trabalho.

Ao professor e orientador Elvio Giasson, pelo apoio, dedicação, paciência e ensinamentos.

Aos professores e funcionários do Departamento de Solos da UFRGS, pelo convívio, especialmente, aos professores da Gênese, Morfologia, Classificação e Levantamento de solos pelas contribuições científicas e experiências profissionais para a realização dos estudos desta tese.

À todos os colegas do PPG Solos da UFRGS. Em especial aos colegas do "aquário" Tatiane Bagatini, Joelma Murliki, Pedro Höfig e Benito Bonfatti.

Enfim, à todos aqueles que, direta ou indiretamente, contribuíram para a realização desta pesquisa.

TESTES METODOLÓGICOS PARA O MAPEAMENTO DIGITAL DE CLASSES DE SOLOS UTILIZANDO ÁRVORES DE DECISÃO[1]/

Autor: Rodrigo Teske
Orientador: Prof. Elvio Giasson

RESUMO

O Mapeamento Digital de Solos (MDS) se utiliza de modelos quantitativos para inferir as variações espaciais e temporais dos solos. Embora venha sendo empregado mundialmente, o MDS ainda não apresenta uma padronização de métodos e materiais. O objetivo deste trabalho foi avaliar e comparar o uso de diferentes metodologias e materiais para análise de dados e predição de ocorrência de classes de solos. Esta pesquisa é composta de uma revisão bibliográfica e de três estudos de predição de ocorrência de classes de solos utilizando técnicas do MDS. Na revisão bibliográfica é discutido e exemplificado o uso de algoritmos de árvores de decisão no MDS, sendo enfatizado o algoritmo CART (*Classification And Regression Tree*). No primeiro estudo, realizado no município de Dois Irmãos, foram avaliados e comparados os efeitos do uso de diferentes modelos digitais de elevação (MDE) sobre a capacidade preditiva dos modelos de predição de ocorrência de classes de solos. Os modelos preditores foram treinados com dados dos atributos do terreno derivados dos diferentes MDE e com informações de solos extraídas do mapa pedológico, na escala de 1:20.000. Os MDE com resolução espacial de 90 m possibilitaram gerar os modelos preditores mais acurados. Na bacia do Rio Santo Cristo, dois estudos foram desenvolvidos. No primeiro estudo realizado nesta bacia foi avaliado e comparado o uso de três esquemas de amostragem de dados para o treinamento dos modelos. A correlação foi gerada com dados de atributos do terreno e informações de solos oriundas de um mapa convencional de solos na escala de 1:50.000. Os esquemas de amostragem influenciaram na acurácia dos modelos preditores, sendo o modelo preditor treinado com dados da amostragem aleatória simples o mais acurado. No segundo estudo realizado para a bacia do Rio Santo Cristo, foi desenvolvido e avaliado um método que concilia o conhecimento pedológico às técnicas do MDS. Primeiramente, foi realizada a predição de ocorrência de classes de solos correlacionando atributos do terreno e a taxonomia de perfis de solos georreferenciados. Esta distribuição espacial das classes de solos foi utilizada para o pedólogo delinear manualmente as unidades de mapeamento de solos (UM). O mapa de UM gerado pelo método proposto apresentou valor da acurácia (avaliada pela verdade de campo) semelhante à de um mapa convencional de solos já existente. Assim, a associação do conhecimento do pedólogo à predição de classes de solo pelas técnicas do MDS demonstrou ser um método especialmente útil na falta de mapas pedológicos de referência para treinamento dos modelos preditores.

[1]/ Tese de Doutorado em Ciência do Solo. Programa de Pós-Graduação em Ciência do Solo, Faculdade de Agronomia, Universidade Federal do Rio Grande do Sul. Porto Alegre. (73 p.) Julho, 2014. Trabalho realizado com apoio financeiro do CNPq e da CAPES.

METHODOLOGICAL TESTS FOR DIGITAL SOIL CLASS MAPPING BY USING DECISION TREES[1]/

Author: Rodrigo Teske
Adviser: Prof. Elvio Giasson

ABSTRACT

The Digital soil mapping (DSM) is used to infer spatial and temporal variations of soil by using quantitative models. Although it has been used worldwide, DSM does not present standardized methods and materials. The objective of this study was to evaluate and compare the use of different methodologies and materials for data analysis and prediction of occurrence of soil class. This research consists of a scientific review and three studies of prediction of occurrence of soil class using DSM techniques. In the scientific review is discussed and exemplified the use of decision tree algorithms to generate predictive models of occurrence of soil class, being emphasized the CART algorithm (Classification And Regression Tree). In the first predictive study, performed in the Dois Irmãos county, were evaluated and compared the effects of using different digital elevation models (DEM) on the ability models to predict the occurrence of soils classes. The prediction models were trained with data from terrain attributes derived from different DEM and soils information extracted from soil map at scale 1:20,000. The DEM with 90 m spatial resolution made it possible to generate the predictive models most accurate. In the Rio Santo Cristo basin, two studies were developed. In the first study of this basin was evaluated and compared three data sampling schemes for training the models. The correlation was generated with data from terrain attributes and soil information derived from a conventional soil map at a scale of 1:50,000. The sampling scheme to influence the accuracy of predictive models, with the model predictor trained with data from simple random sampling was more accurate. In the second study for the Santo Cristo river basin, was developed and evaluated a method that reconciles the soil knowledge of MDS techniques. First, was performed the prediction of occurrence of soil types correlating terrain attributes and taxonomy of soil profiles georeferenced. This spatial distribution of soil classes was used by the pedologist for delineating soil mapping units (MU). The map of MU generated by the proposed method showed values of accuracy (assessed by ground truth) similar to that of a conventional map of existing soils. Therefore, the combination of the knowledge of pedologist the prediction of soil classes by MDS techniques proved to be a particularly useful method in the absence of soil maps of reference for training predictive models.

[1]/ Doctoral thesis in Soil Science, Graduate Program in Soil Science, Faculty of Agronomy, Federal University of Rio Grande do Sul, Porto Alegre. (73 p.) July, 2014. Research supported by CNPq and CAPES.

SUMÁRIO

RELAÇÃO DE TABELAS

Página

RELAÇÃO DE FIGURAS

Página

RELAÇÃO DE ABREVIATURAS E SÍMBOLOS

ACL	Árvore de Classificação
AD	Árvore de Decisão
ADr	Árvore de Decisão Reforçada
AF	Aerofotogrametria
AG	Acurácia Geral
AM	Acurácia do Mapeador
AR	Árvore de Regressão
ArcGis	Programa computacional de sistema de informação geográfica
ASTER	Advanced Spaceborne Thermal Emission and Reflection Radiometer
CART	Classification And Regression Tree
Cfa	Clima subtropical úmido quente
Cfb	Clima subtropical úmido temperado
CN	Curva de Nível
COS	Carbono Orgânico do Solo
CS	Classe de Solo
DP	Divisão Percentual
ESRI	*Environmental Systems Research Institute*
GPS	Sistema de Posicionamento Global
IBGE	Instituto Brasileiro de Geografia e Estatística
M	Número de elementos no nó final
MDE	Modelos Digitais de Elevação
MDS	Mapeamento Digital de Solos
RS	Rio Grande do Sul
SiBCS	Sistema Brasileiro de Classificação de Solos
SIG	Sistemas de Informação Geográfica
SimpleCart	Algoritmo de árvore de decisão
SR	Sensores Remotos
SRTM	*Shuttle Radar Topographic Mission*
TDIDT	*Top-Down Induction of Decision Tree*
UM	Unidade de Mapeamento de Solos
VA	Validação Aparente
VC	Validação Cruzada
WEKA	Waikato Environment for Knowledge Analysis

1. INTRODUÇÃO GERAL

Um levantamento pedológico consiste não somente em mapear as unidades de mapeamento de solo (UM), mas também, de fornecer informações do ambiente em que cada tipo de solo se encontra. A disponibilidade destas informações, em escala adequada, é de fundamental importância para o planejamento de atividades relacionadas ao uso e manejo de solos. Com o aumento da população humana, além da utilização do solo para fins urbanos e industriais como descartes de resíduos e construções civis, mais áreas virão a ser destinadas à agricultura para suprir as demandas de produtos agrícolas como alimentos, combustível e madeira.

As informações de solo existentes no Brasil são, em sua maioria, oriundos de levantamentos Exploratórios e de Reconhecimento. O Rio Grande do Sul possui seu território coberto por mapa de solos na escala de 1:1.000.000 e na escala de 1:750.000. Mapas em escalas iguais ou maiores que 1:100.000 existem somente em algumas áreas do estado. A escassez de mapas de solos se deve ao fato de os levantamentos convencionais demandarem altos custos e muito tempo para realizar as suas diversas etapas. Como uma alternativa para auxiliar a execução dos mapeamentos de solos, vem sendo testado o uso do Mapeamento Digital de Solos (MDS). O MDS tem sido impulsionado pelos avanços tecnológicos como o Sistema de Posicionamento Global (GPS), imagens de sensoriamento remoto, Modelos Digitais de Elevação (MDE), Sistemas de Informação Geográfica (SIG) e os algoritmos de aprendizagem de máquina.

As técnicas do MDS utilizadas para predizer a ocorrência das classes de solo também visam estabelecer relações entre os solos e as características da paisagem. Todavia, no MDS estas relações solo-paisagem são realizadas de forma mais quantitativa, pelo uso de modelos que correlacionam variáveis ambientais com a ocorrência espacial dos solos. Dentre as variáveis ambientais, os atributos do terreno derivados de MDE são gerados computacionalmente e representam condições ambientais que influenciam diversos processos atuantes na formação do solo, como a velocidade do fluxo superficial e subsuperficial de água, o teor de água no solo, o potencial dos processos erosivos e deposicionais, entre outros.

Os modelos preditores e mapas de ocorrência de solos gerados pelas técnicas do MDS, à semelhança dos mapeamentos convencionais de solo, apresentam erros que devem ser identificados e quantificados. A comparação (pixel a pixel) do mapa gerado com um mapa de solo já existente é um método comumente empregado no MDS. A partir das informações confrontadas, uma matriz de confusão é gerada e os dados de referência são dispostos em colunas e os dados das classes preditas são alocados nas linhas. Assim, a identificação das classes mais concordantes (diagonal principal) e das classes com mais erros de classificação (nas demais células) podem ser observadas diretamente na matriz de confusão. A quantificação da concordância e dos erros de classificação pode ser obtida por indicadores comumente denominados de acurácia geral, acurácia do mapeador, acurácia do usuário, erros de inclusão e erros de omissão. Todavia, é recomendado que a avaliação da acurácia de mapas de solos, tanto os convencionais como os digitais, seja feita pela verdade de campo, a fim de se obter os erros e acertos dos mapas com base na real distribuição dos solos na paisagem.

Com o intuito de desenvolver possíveis contribuições para a padronização de métodos e formulação de protocolos de procedimentos para o MDS foram desenvolvidos quatro estudos, consistindo de uma revisão bibliográfica e três estudos utilizando técnicas do MDS para duas regiões do Rio Grande do Sul. Na Revisão Bibliográfica do Capítulo I é apresentado e discutido o uso de árvores de decisão para treinamento e geração de modelos preditores, sendo exemplificados procedimentos matemáticos realizados por algoritmos de árvores de classificação durante a construção de um modelo

preditor de ocorrência de classe de solo. Nesta revisão é apresentado o estado da arte do uso destes algoritmos no MDS, as vantagens e perspectivas da utilização das árvores de decisão para trabalhos de predição de ocorrência de solos que se utilizam de variáveis correlacionadas com os solos.

No Capítulo II são apresentadas as avaliações e comparações do uso de diferentes MDE (tipos e resolução espacial) sobre a capacidade preditiva de modelos preditores de ocorrência de UM que foram gerados por árvores de classificação. A área deste estudo foi o município de Dois Irmãos (RS), que apresenta predomínio de relevo plano e suave ondulado. Isto possibilitou gerar modelos preditores de ocorrência de classes de solos com diferentes valores da acurácia e quantidades de UM preditas, conforme o MDE utilizado para gerar as variáveis correlacionadas (atributos do terreno).

Os estudos realizados na microbacia do Rio Santo Cristo (RS) estão apresentados e discutidos nos Capítulos III e IV. Os estudos do Capítulo III se referem a avaliações do uso de diferentes esquemas de amostragem dos dados correlacionados (solos e atributos do terreno) usados para o treinamento dos modelos preditores, os quais foram usados para gerar os mapas digitais de classes de solos. Adicionalmente, a capacidade preditiva dos modelos gerados foi estimada por diferentes métodos de validação. Os resultados demonstraram que a seleção do esquema de amostragem pode influenciar na capacidade preditiva dos modelos e, dependendo do método de avaliação utilizado, a acurácia dos modelos pode ser superestimada.

O Capítulo IV apresenta um método especialmente útil para a predição de ocorrência de classes e de unidades de mapeamento de solos na falta de mapas pedológicos de referência para o treinamento de modelos. Ao utilizar a predição de ocorrência de tipos de solos, a partir de perfis de solos georreferenciados, com delineamento das UM realizado manualmente, demonstrou que as técnicas preditivas do MDS associadas ao conhecimento e experiência de um pedólogo podem ser empregadas simultaneamente para a elaboração de mapas digitais de solos.

2. CAPÍTULO I – ÁRVORES DE DECISÃO E SEU USO NO MAPEAMENTO DIGITAL DE SOLOS

2.1 Introdução

Como uma alternativa para auxiliar a execução dos levantamentos convencionais, o mapeamento digital de solos (MDS) tem sido mundialmente utilizado para suprir informações pedológicas (McBratney et al., 2003). Os trabalhos com MDS utilizam modelos quantitativos para inferir as variações espaciais e temporais dos solos, a partir de observações a campo, do conhecimento pedológico e de variáveis ambientais correlacionadas com os solos (Lagacherie & McBratney, 2007). Segundo Omuto et al. (2013), diversos métodos de predição podem ser usados e são classificados em três grandes grupos: os geoestatísticos, os não geoestatísticos e os métodos mistos. A utilização destes métodos varia conforme o objetivo do projeto e das informações disponíveis a respeito dos solos, material de origem, vegetação e relevo, entre outros. Embora o MDS ainda não possua um protocolo de metodologias padronizadas (Omuto et al., 2013), a predição espacial de classes de solo e unidades de mapeamento tem sido realizada com uso de técnicas preditivas como a lógica fuzzy, regressão logística e árvores de decisão (Zhu et al., 2001; ten Caten et al., 2011; Giasson et al., 2013).

As técnicas preditivas são comumente utilizadas em planejamentos estratégicos de setores privados e governamentais, como áreas da saúde,

mercado de valores, estratégias de consumo no varejo e atacado, defesa e serviços de transporte, entre outros (Witten et al., 2011; Rokach & Maimon, 2008). Contudo, na área da Ciência do Solo a utilização destas técnicas preditivas é ainda muito incipiente, principalmente no Brasil, dado ao recente avanço nas áreas de informática e tecnologia, bem como à falta de investimento em treinamento de pedólogos e execução de levantamentos de solo no país (Mendonça-Santos & Santos, 2006). Por conta disso, ainda há poucos pesquisadores com conhecimento adequado e completo nas diversas áreas necessárias ao emprego das técnicas do MDS, tais como modelagem, estatística, pedologia, cartografia, gestão de banco de dados e integração destes aos programas estatísticos, de mineração de dados e de sistemas de informação geográfica, que necessitam de constante aprendizado e treinamento (Omuto et al., 2013).

Como todo assunto tecnológico e inovador é visto de forma mais criteriosa e cética, as técnicas preditivas são, muitas vezes, utilizadas para a predição de ocorrência de solos sem, necessariamente, entender os procedimentos matemáticos envolvidos ou avaliar como as variáveis (solos e ambiente) são relacionadas na construção do modelo. Assim, bibliografias científicas contendo os procedimentos de construção de modelos preditores se tornam necessários para auxiliar e melhorar o entendimento a respeito destas técnicas preditivas para o MDS. O objetivo deste trabalho foi apresentar uma revisão bibliográfica com procedimentos matemáticos usados por algoritmos de árvore de decisão, as vantagens, as aplicações e perspectivas do uso de árvores de decisão em trabalhos com MDS.

2.2 Desenvolvimento

Os algoritmos de mineração de dados, em sua maioria, são baseados na aprendizagem indutiva, em que um modelo é construído a partir de um número suficiente de exemplos ou dados para treinamento (Rokach & Maimon, 2008). Basicamente, são quatro estilos diferentes de aprendizagem em mineração de dados: i) a aprendizagem de classificação, em que o treinamento do algoritmo é realizado a partir de um conjunto de exemplos; ii) a aprendizagem de associação, em que qualquer associação entre as

características é procurado, e não apenas aqueles que predizem um valor ou determinada classe; iii) em *clusters*, em que são procurados grupos de exemplos semelhantes; iv) a predição numérica, em que o resultado a ser predito não é uma variável categórica discreta (classe), mas sim variáveis numéricas (Witten et al., 2011).

O aprendizado indutivo pode ser dividido em método não supervisionado e método de aprendizagem supervisionado (Chapelle et al., 2006). O processo de aprendizagem não supervisionada ocorre por meio de observação e descoberta, ou seja, não existem dados classificados de exemplos e, neste caso, o método tem que encontrar atributos estatísticos relevantes para a predição. Neste conceito se enquadram os algoritmos de análise de grupos (*clustering*), os modelos de mistura e as redes neuronais não supervisionadas. Na aprendizagem supervisionada, cada atributo de entrada (variável preditora ou independente) está acompanhado de uma classe ou de um valor alvo (variável resposta ou dependente). Neste conceito, se incluem as técnicas de estatística multivariada como a regressão, análise discriminante, a regressão logística e as técnicas de reconhecimento de padrões como as redes neurais supervisionadas e as árvores de decisão (regressão e classificação) (Rokach & Maimon, 2008; Basgalupp, 2010).

Em mineração de dados, uma árvore de decisão é um modelo preditivo que pode ser usado para representar tanto as classificações como os modelos de regressão. Quando uma árvore de decisão é usada para tarefas de predição categórica (aprendizagem de classificação), é mais apropriado chamá-la de árvore de classificação, e quando usada para tarefas de predição numérica é denominada de árvore de regressão (Rokach & Maimon, 2008). Nesta revisão, serão apresentados e discutidos algoritmos e modelos de árvores de classificação. Uma abordagem mais detalhada com árvores de regressão pode ser encontrada em Yohannes & Webb (1999).

A construção das árvores de decisão é realizada conforme o conjunto de dados disponibilizado para treinamento do algoritmo. As árvores de decisão têm a função de particionar recursivamente este conjunto de dados até que cada subconjunto obtido contenha casos de uma única classe ou valor. Os modelos de árvores de decisão são apresentados em uma estrutura hierárquica que se desenvolvem da raiz para as folhas. Assim, a representação

hierárquica é traduzida em uma progressão da análise de dados para desempenhar uma tarefa de predição, tornando as árvores de decisão auto-explicativas e sem a necessidade de se tornar um especialista em mineração de dados para utilizá-las (Rokach & Maimon, 2008). Desta forma, a construção das árvores de decisão é regida pelo princípio “dividir para conquistar”, em que cada nível de uma árvore, um problema mais complexo de classificação é decomposto em subproblemas mais simples, resultando, assim, na geração de nós descendentes, em que a heterogeneidade da variável resposta a ser predita é mais atenuada (Tan et al., 2005). Desta forma, uma árvore de decisão é constituída de um nó “raiz”, de nós “internos” ou “de teste” que apresentam a indicação de qual variável ou de quais valores das variáveis preditoras foram divididos, e os nós que apresentam a predição da variável resposta, que são chamados de "folhas" e também de “nó terminal" ou “nó decisão". Na Figura 1 o nó raiz é representado por um círculo, os nós internos são representados como retângulos e as folhas são indicados como triângulos.

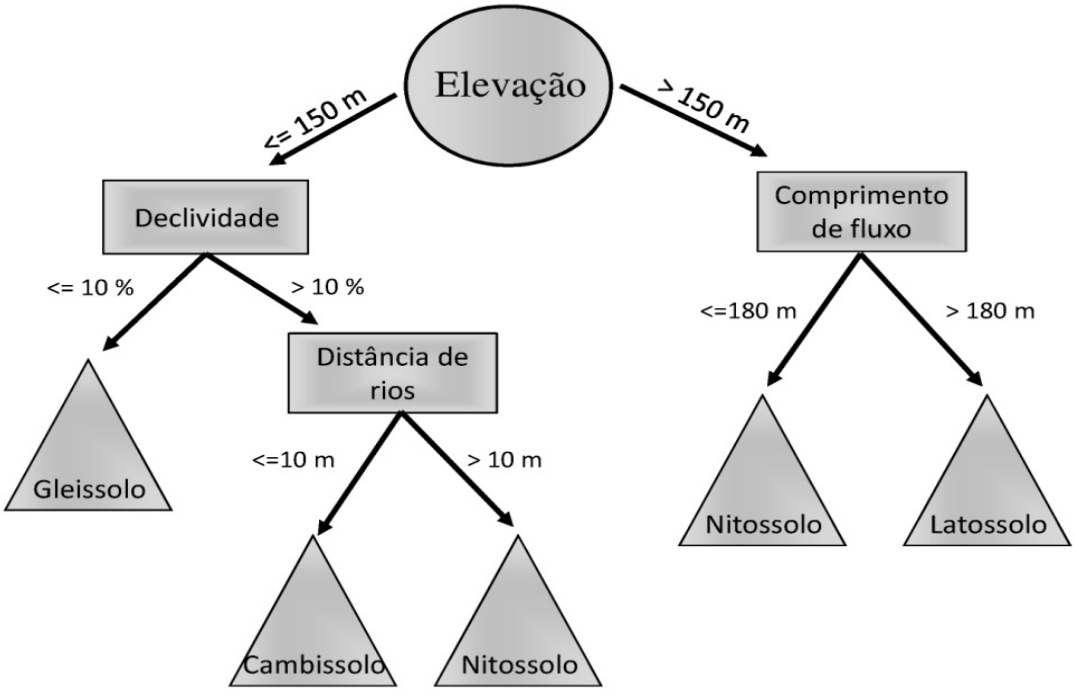

Figura 1. Exemplo de uma árvore de classificação na predição de classes de solo.

Há muitas maneiras de uma árvore de decisão ser estruturada a partir de um conjunto de dados. O *Top-Down Induction of Decision Tree* (TDIDT) é um algoritmo utilizado como base para muitos algoritmos de indução

de árvores de decisão, dentre eles os mais conhecidos como ID3 (Quinlan, 1986), C4.5 (Quinlan, 1993) e CART (Breiman et al., 1984). O TDIDT é um algoritmo recursivo de busca, que procura sobre um conjunto de variáveis preditoras, aqueles que melhor dividem os dados em subconjuntos. Inicialmente, todas as variáveis preditoras são colocadas em um nó raiz. Em seguida, uma variável preditora é selecionada para representar o teste desse nó e, assim, dividir os exemplos em subconjuntos de exemplos (classes ou valores). A partir da raiz, o resultado obtido a partir de um teste de decisão é determinado em cada nó interno, dividindo o conjunto de dados em subárvores, e o processo de teste se inicia pela raiz da subárvore seguinte. Esse processo, conhecido como particionamento recursivo se repete até que todas as variáveis resposta tenham sido classificadas ou até que todas as variáveis preditoras tenham sido utilizadas (Basgalupp, 2010; Rokach & Maimon, 2008). O termo recursivo é usado para indicar que cada nó “filho”, por sua vez, se tornará um nó “pai”, a menos que seja um nó terminal (Moisen, 2008).

Na construção da árvore de decisão, procura-se associar a cada nó interno a variável preditora mais informativa. Todavia, cada algoritmo se utiliza de diferentes métodos para selecionar essa variável mais informativa, o que por sua vez, irá influenciar na topologia, no tamanho da árvore e na acurácia do modelo preditor gerado. Segundo (Loh, 2011), ao utilizar um algoritmo de árvore de decisão devem ser atendidos pelo menos dois objetivos principais: (i) particionar os dados em cada etapa e (ii) parar o particionamento. A maioria dos algoritmos utiliza divisões binárias, onde a variável preditora e o ponto de divisão são frequentemente encontrados por uma pesquisa exaustiva e são escolhidos para minimizar a impureza de cada nó (para árvore de classificação) ou a soma dos quadrados dos resíduos (para árvore de regressão), conhecidos como critérios de divisão (Moisen, 2008). Em relação ao segundo objetivo, cada árvore continua a crescer tanto quanto for possível para que cada subárvore obtida contenha casos de uma única classe (ou valor), ou até certo tamanho cuja definição é realizada pelo usuário utilizando regras de parada de divisão ou podas de árvores (Timofeev, 2004; Loh, 2011).

A partição recursiva binária é aplicada tanto nas árvores de classificação como nas de regressão. Durante a construção das árvores há duas principais operações (i) a avaliação das divisões para cada variável

preditora e seleção da melhor divisão e (ii) a criação de partições usando a melhor divisão. Tendo determinado o melhor desdobramento total, subdivisões podem ser criadas por um critério de divisão para os dados. A complexidade está em determinar a melhor divisão para cada variável preditora. Para isto, um índice de divisão é usado para avaliar a qualidade das divisões, sendo este critério adotado, diferente para cada tipo de árvore de decisão (Rokach & Maimon, 2008). Desta forma, a seleção de uma variável e um ponto de corte nos valores dessa variável é realizada para maximizar uma medida de entropia dos nós "filhos" relativamente ao nó "pai" (C4.5 e ID3) ou para minimizar uma medida de impureza (CART) (Yohannes & Webb, 1999). Segundo diversos autores (Timofeev, 2004; Moisen, 2008; Loh, 2011), dois destes critérios são mais comumente utilizados em problemas de classificação: o Índice Gini no algoritmo CART e o ganho de informação (Entropia) no algoritmo J48.

O algoritmo J48 é uma implementação em Java do algoritmo C4.5 (Quinlan, 1993) no programa Weka (Hall et al., 2009), que gera árvores de classificação a partir de um conjunto de dados de treinamento e em cada nó, o algoritmo escolhe um atributo que mais eficientemente divida o conjunto das amostras em subconjuntos homogêneos. Em cada nó, as variáveis preditoras disponíveis são separadas em função das classes de exemplo e a subdivisão dos dados é realizada quando se obtém a menor medida de impureza calculada pela entropia (Eq. 1) e maior ganho de informação (Eq. 2). Determinando assim, o quão boa é a condição de teste realizada, ao subtrair o grau de entropia dos nós-filhos (após a divisão) da medida de entropia do nó-pai (antes da divisão) (Quinlan, 1993; Du & Zhan, 2002).

$$\mathbf{Entropia(nó)} = -\sum_{i=1}^{c} \mathbf{p}\left(\frac{\mathbf{i}}{\mathbf{nó}}\right) * \mathbf{log_2}\left[\mathbf{p}\left(\frac{\mathbf{i}}{\mathbf{nó}}\right)\right] \qquad \text{(Eq. 1)}$$

Onde $\mathbf{p}\left(\frac{\mathbf{i}}{\mathbf{nó}}\right)$ é a fração dos registros pertencentes à classe **i** no nó, e **c** é o número de classes.

$$\mathbf{Ganho_{entropia}(E,G)} = \mathbf{Entropia(p)} - \sum_{f=1}^{n} \left(\frac{|\mathbf{Nf}|}{|\mathbf{Np}|} * \mathbf{Entropia(f)}\right) \qquad \text{(Eq. 2)}$$

onde **n** é o número de nós-filhos, **Np** é o número total de objetos do nó-pai e **Nf** é o número de exemplos associados ao nó-filho **f**.

O algoritmo SimpleCart é uma implementação em Java do algoritmo CART baseado em Breiman et al. (1984). CART é uma metodologia estatística não paramétrica desenvolvida para analisar questões de classificação e de regressão. As árvores de decisão indutivas geradas pelo algoritmo CART são construídas pela aplicação de divisão em subárvores mais homogêneas usando o critério de Gini (Eq. 3) (Yohannes & Webb, 1999; Loh, 2011).

$$\textbf{Índice}_{\textbf{Gini}}(\textbf{nó}) = 1 - \sum_{i=1}^{c} \left[p\left(\frac{i}{\textbf{nó}}\right) \right]^2 \qquad \text{(Eq.3)}$$

onde $p\left(\frac{i}{nó}\right)$ é a fração dos registros pertencentes à classe **i** no nó, e **c** é o número de classes.

Assim como no cálculo do ganho de informação com a entropia, o ganho de informação com o Índice Gini é calculado pela diferença entre o índice Gini antes e após a divisão. Essa diferença é calculada pela Eq. 4.

$$\textbf{Ganho}_{\textbf{Gini}} = \textbf{Índice}_{\textbf{Gini}}(\textbf{pai}) - \sum_{f=1}^{n} \left[\left(\frac{Nf}{Np}\right) * \textbf{Índice}_{\textbf{Gini}}(\textbf{f}) \right] \qquad \text{(Eq.4)}$$

onde **n** é o número de valores de nós-filhos, **Np** é o número total de objetos do nó-pai e **Nf** é o número de exemplos associados ao nó-filho **f**.

A fim de exemplificar, a seguir são demonstradas etapas da divisão dos dados usando o critério de índice Gini para a construção de uma árvore de classificação. Para isto, foram utilizadas informações adaptadas de Tan et al. (2005). Estas informações são hipoteticamente referentes a 20 pontos amostrais que estão relacionados com a Vegetação Primária (campo ou floresta), com a Fase de Relevo (plano, ondulado e montanhoso), com a altitude do local amostrado em metros e com as Classes de Solos (A e B) a que pertencem cada um dos pontos amostrais.

2.3 Dados categóricos

A variável Classe de Solo possui o mesmo número de registros em cada classe. Assim, antes de qualquer divisão cada classe apresenta uma distribuição de 50% (Tabela 1 e Figura 2). Isto pode ser constatado pelo emprego do índice Gini (Eq. 5), com uma distribuição de 0,50 para classe A e 0,50 para classe B.

Tabela 1. Dados de exemplos hipotéticos para partição binária usando o índice Gini (adaptado de Tan et al., 2005).

Ponto amostral	Vegetação primária	Elevação (m)	Fase do relevo	Classe de solo
1	Campo	180	Plano	A
2	Campo	176	Ondulado	A
3	Campo	172	Ondulado	A
4	Campo	167	Ondulado	A
5	Campo	178	Ondulado	A
6	Campo	181	Ondulado	A
7	Floresta	167	Ondulado	A
8	Floresta	168	Ondulado	A
9	Floresta	170	Ondulado	A
10	Floresta	173	Montanhoso	A
11	Campo	185	Plano	B
12	Campo	186	Plano	B
13	Campo	190	Plano	B
14	Campo	185	Montanhoso	B
15	Floresta	170	Montanhoso	B
16	Floresta	165	Montanhoso	B
17	Floresta	170	Montanhoso	B
18	Floresta	172	Montanhoso	B
19	Floresta	165	Montanhoso	B
20	Floresta	168	Montanhoso	B

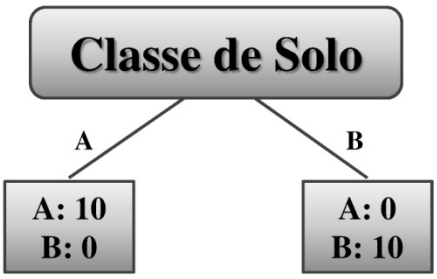

Figura 2. Divisão binária dos dados pela classe de solos.

$$\mathbf{Gini_{AB}} = 1-\left(\frac{10}{20}\right)^2 - \left(\frac{10}{20}\right)^2 = 0{,}50 \qquad \text{(Eq.5)}$$

Caso a divisão dos dados seja realizada utilizando a variável Vegetação Primária como raiz (Figura 3), os nós-filhos apresentarão uma distribuição de (0,60 e 0,40) para a vegetação de campo e de (0,40 e 0,60) para a vegetação de floresta. Cada um desses nós filhos terão um índice Gini de 0,48 (Eq. 6 e Eq. 7).

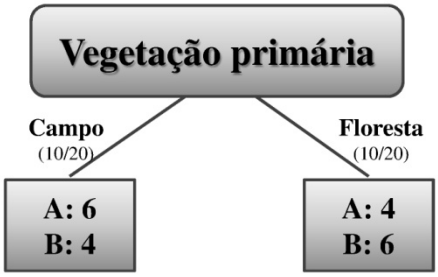

Figura 3. Divisão binária dos dados pela vegetação primária.

$$\mathbf{Gini_{Campo}} = 1-\left(\frac{6}{10}\right)^2 - \left(\frac{4}{10}\right)^2 = 0{,}48 \qquad \text{(Eq.6)}$$

$$\mathbf{Gini_{Floresta}} = 1-\left(\frac{4}{10}\right)^2 - \left(\frac{6}{10}\right)^2 = 0{,}48 \qquad \text{(Eq.7)}$$

A média ponderada do índice de Gini para estes nós descendentes é igual a 0,48 (Eq. 8).

$$\mathbf{Gini_{CampoFloresta}} = \left[\left(\frac{10}{20}\right) \text{ x } 0{,}48\right] + \left[\left(\frac{10}{20}\right) \text{ x } 0{,}48\right] = 0{,}48 \qquad \text{(Eq.8)}$$

A distribuição dos dados calculada pelo índice Gini varia em uma escala de 0 a 1 e indica como é a distribuição dos dados. Quanto mais próximo de 1 mais heterogênea e quanto mais próximo de 0 esta distribuição é mais homogênea. Como o conjunto de dados apresenta mais uma variável preditora Fases de Relevo, isto permite realizar outras divisões visando encontrar subdivisões mais homogêneas. Desta forma, ao dividir os dados em um grupo binário usando como nó raiz Fases de Relevo (Figura 4), o índice Gini do nó contendo o grupamento {Ondulado e Montanhoso} foi de 0,49 e o índice Gini do nó {Plano} foi de 0,38 (Eq. 9 e Eq. 10). A média ponderada do índice de Gini para estes grupamentos da Fase do Relevo I foi igual a 0,47 (Eq. 11).

$$\mathbf{Gini_{Ondulado\_Montanhoso}} = 1-\left(\frac{9}{16}\right)^2 - \left(\frac{7}{16}\right)^2 = 0{,}49 \qquad \text{(Eq.9)}$$

$$\mathbf{Gini_{Plano}} = 1-\left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2 = 0{,}38 \qquad \text{(Eq.10)}$$

$$\mathbf{Gini_{FasedoRelevo\_I}} = \left[\left(\frac{16}{20}\right) \text{ x } 0{,}4922\right] + \left[\left(\frac{4}{20}\right) \text{ x } 0{,}375\right] = 0{,}47 \qquad \text{(Eq.11)}$$

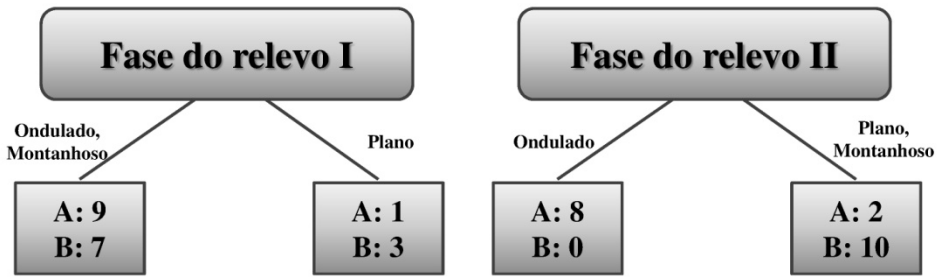

Figura 4. Divisão binária dos dados por fases do relevo.

De forma semelhante, a segunda combinação binária de grupamentos de Fase do Relevo II {Ondulado} e {Plano e Montanhoso} resultou valores de índice Gini iguais a 0 e 0,28, respectivamente (Eq. 12 e Eq. 13). A média ponderada do índice de Gini foi de 0,17 (Eq. 14).

$$\mathbf{Gini_{Ondulado}} = 1-\left(\frac{8}{8}\right)^2 - \left(\frac{0}{8}\right)^2 = 0 \qquad \text{(Eq.12)}$$

$$\mathbf{Gini_{Plano\_Montanhoso}} = 1-\left(\frac{2}{12}\right)^2 - \left(\frac{10}{12}\right)^2 = 0{,}28 \qquad \text{(Eq.13)}$$

$$\mathbf{Gini_{FasedoRelevo\_II}} = \left[\left(\frac{8}{20}\right) \text{ x } 0\right] + \left[\left(\frac{12}{20}\right) \text{ x } 0{,}2777\right] = 0{,}17 \qquad \text{(Eq.14)}$$

A subdivisão dos dados usando grupamentos binários com Fases do Relevo, especialmente o grupamento binário Fase do Relevo II ({Ondulado} e {Plano e Montanhoso}) resultou na partição mais homogênea dos dados. Assim, a constituição da árvore é representada com o nó raiz Fases do Relevo, e as subárvores são construídas a partir de grupamentos binários que também venham a apresentar uma maior homogeneidade da distribuição dos dados, ora avaliadas pelo índice Gini na CART, ora avaliada pela entropia na J48 tal como demonstrado mais detalhadamente em Quinlan (1986); Du & Zhan (2002) e Witten et al. (2011).

2.4 Dados numéricos

Quando as variáveis preditoras se apresentarem como dados numéricos, a condição de teste pode ser expressa de duas maneiras: i) como um teste de comparação com resultados binários ($Vp<x$) ou ($Vp\geq x$), onde Vp é a variável preditora e x é o valor de teste para o particionamento recursivo

(ponto de referência); ii) como uma consulta do intervalo dos dados. A divisão binária dos dados é, geralmente, a mais utilizada e é realizada através da escolha de um ponto intermediário entre dois valores diferentes e consecutivos, após ordenação dos dados por ordem crescente dos valores da variável numérica. Assim, o algoritmo de árvore de decisão deve considerar todos os possíveis pontos de divisão em x, e selecionar aquele que produz a melhor partição, ou seja, a mais homogênea. Como exemplo, os dados da Elevação da Tabela 1 foram colocados em ordem crescente (Tabela 2).

Tabela 2. Dados de exemplos hipotéticos e postos em ordem crescente de altura. (adaptado de Tan et al., 2005).

Ponto amostral	Vegetação primária	Elevação (m)	Fase do relevo	Classe de solo
16	Floresta	165	Montanhoso	B
19	Floresta	165	Montanhoso	B
4	Campo	167	Ondulado	A
7	Floresta	167	Ondulado	A
8	Floresta	168	Ondulado	A
20	Floresta	168	Montanhoso	B
9	Floresta	170	Ondulado	A
15	Floresta	170	Montanhoso	B
17	Floresta	170	Montanhoso	B
3	Campo	172	Ondulado	A
18	Floresta	172	Montanhoso	B
10	Floresta	173	Montanhoso	A
2	Campo	176	Ondulado	A
5	Campo	178	Ondulado	A
1	Campo	180	Plano	A
6	Campo	181	Ondulado	A
11	Campo	185	Plano	B
14	Campo	185	Montanhoso	B
12	Campo	186	Plano	B
13	Campo	190	Plano	B

Diversas possibilidades de divisão dos dados podem ser realizadas pelo algoritmo em computadores. De forma resumida, são apresentados os resultados de duas subdivisões dos dados. O ponto de referência (Elevação 1) está entre os valores de elevação 168 m e 170 m (169 m) e apresentou uma distribuição de 6 exemplos de Classes de Solos para elevações menores ou iguais a 169 m (3 classes A e 3 classes B) e para elevações maiores que 169 m foram 14 exemplos (7 classes A e 7 classes B). Quando realizada a divisão dos dados com ponto de referência de 183 m (Elevação 2), aos dados foram

divididos em 16 exemplos (10 A e 6 B) quando menores ou iguais a 183 m, e quando maiores a 183 m apresentaram 4 exemplos da classe B (Figura 5).

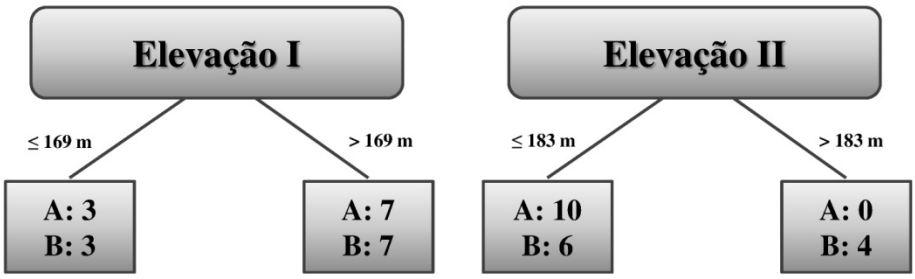

Figura 5. Divisão binária dos dados por elevação (m).

Ao utilizar o ponto de referência igual a 169 m a distribuição dos dados foi de 0,50 para a divisão ≤ 169 m e de 0,50 para a divisão binária > 169 m (Eq. 15 e Eq. 16). Logo, a média ponderada do índice Gini para estes nós descendentes é de 0,50.

$$\mathbf{Gini_{\leq 169\ m}} = 1-\left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2 = 0{,}50 \qquad \text{(Eq.15)}$$

$$\mathbf{Gini_{>169\ m}} = 1-\left(\frac{7}{14}\right)^2 - \left(\frac{7}{14}\right)^2 = 0{,}50 \qquad \text{(Eq.16)}$$

Ao utilizar o ponto de referência igual a 183 m a distribuição dos dados foi de 0, 47 para a divisão ≤ 183 m e de 0 para a divisão binária > 183 m (Eq. 17 e Eq. 18). Logo, a média ponderada do índice Gini para estes nós descendentes é de 0,38 (Eq. 19).

$$\mathbf{Gini_{\leq 183\ m}} = 1-\left(\frac{10}{16}\right)^2 - \left(\frac{6}{16}\right)^2 = 0{,}47 \qquad \text{(Eq.17)}$$

$$\mathbf{Gini_{>183\ m}} = 1-\left(\frac{4}{4}\right)^2 - \left(\frac{0}{4}\right)^2 = 0 \qquad \text{(Eq.18)}$$

$$\mathbf{Gini_{Elevação\ 2}} = \left[\left(\frac{16}{20}\right) \text{ x } 0{,}469\right] + \left[\left(\frac{4}{20}\right) \text{ x } 0\right] = 0{,}38 \qquad \text{(Eq.19)}$$

Com isso, a subdivisão dos dados usando o ponto de referência 183 m (Elevação 2), resultou em uma partição mais homogênea do que a partição usando o ponto de referência 169 m (Elevação 1). Todas as demais

possibilidades de subdivisão dos dados são testadas avaliando qual ponto de referencia resulta em partições de dados mais homogênea. Assim, este procedimento é repetido até que os valores de índice de Gini sejam calculados para todos os casos, selecionando a melhor posição de divisão corresponde ao que produz o menor índice de Gini. Quando são utilizados dados numéricos contínuos, a metodologia de cálculo é similar, porém os dados sofrem uma discretização e, após este procedimento, é atribuído um novo valor para cada intervalo de dados discretos, como pode ser observado em Tan et al. (2005).

As árvores de decisão apresentam diversas vantagens como uma ferramenta de classificação (Yohannes & Webb, 1999; Breiman et al., 1984; Witten et al., 2011; Rokach & Maimon, 2008). Dentre as quais, i) as árvores de decisão podem lidar com variáveis preditoras tanto nominais como numéricas; ii) os conjuntos de dados podem conter erros e valores discrepantes, os quais serão alocados em nós-pai ou nós-filhos separadamente sem comprometer a classificação; iii) as árvores de decisão são considerados um método não paramétrico, ou seja, não são realizadas quaisquer suposições da distribuição das variáveis dependentes e independentes e a estrutura do classificador não é afetada por valores discrepantes, colinearidades, heterosedasticidade ou estruturas de distribuição do erro que afetam os procedimentos paramétricos; iv) o resultado da árvore não varia sob uma transformação matemática (logarítmica, raiz quadrada, etc.) das variáveis independentes; v) as árvores de decisão são autoexplicativas e sua compreensão se torna mais fácil ao serem transformadas em um conjunto de regras de classificação (Quinlan, 1986).

Utilizando a árvore de decisão apresentada na Figura 1, podemos transformá-la em um conjunto de proposições ou regras de classificação:

Se elevação <= 150 m e declive <= 10% então, Gleissolo;

Se elevação <= 150 m e declive > 10% e distância de rios <= 10 m então, Cambissolo.

A partir das regras de classificação, as correlações geradas entre a ocorrência espacial dos solos e as variáveis preditoras, tais como as características do relevo e geologia, podem ser identificadas e interpretadas pelo pedólogo. Os modelos preditores mais acurados e que geram as melhores correlações entre as variáveis preditoras e as variáveis preditas são, geralmente, os que apresentam as maiores e mais complexas árvores de

decisão (Breiman et al., 1984; Quinlan, 1986). Todavia, analisar a totalidade das regras e correlações entre as variáveis ambientais e os solos pode não ser vantajoso para as aplicações práticas do MDS, principalmente, quando as árvores se tornarem muito grandes e complexas. Para obter os mapas digitais de solos, as regras de classificação são implementadas em programas de sistemas de informação geográfica (SIG), o que permitem observar e interpretar a distribuição espacial dos solos de acordo com as características da paisagem.

Desta forma, as árvores de decisão têm sido utilizadas mundialmente em trabalhos com MDS (Tabela 3) por serem consideradas robustas, serem capazes de processar grandes volumes de dados sem a necessidade de realizar quaisquer suposições sobre a distribuição dos dados (solos) e as variáveis preditoras (p.e. atributos do terreno), e da não necessidade de realizar transformações dos dados (Witten et al., 2011; Rokach & Maimon, 2008). Além disso, permitem relacionar uma gama de variáveis discretas e contínuas em diferentes escalas e diferentes tipos de informações como as imagens de sensoriamento remoto, os modelos digitais de elevação e seus produtos derivados e os mapas de solos já existentes, entre outros (Scull et al., 2005).

Segundo Henderson et al. (2005), as árvores de decisão podem ser consideradas como ferramentas eficientes para predição de ocorrência de solos por permitirem extrair o conhecimento pedológico a partir de informações pedológicas já existentes e correlacioná-las com a paisagem. Desta forma, uma árvore de decisão pode ser considerada análoga às correlações mentais realizadas pelo pedólogo, ao relacionar as características da paisagem e a distribuição dos solos, dada sua semelhança ao aprendizado humano indutivo (exemplificado), porém, de uma forma mais quantitativa, o que possibilita a sua replicação, diferentemente dos levantamentos convencionais de solos que dificilmente apresentam de forma clara as correlações entre os solos e as características da paisagem (Hartemink et al., 2010; Jafari et al., 2012). Por isso, as árvores de decisão têm sido utilizadas em maior número em trabalhos que envolvam a predição categórica de classes de solo e unidades de mapeamento de solo (Grinand et al., 2008; Minasny & McBratney, 2007), conforme listados na Tabela 3.

Tabela 3. Relação de publicações em mapeamento digital de solos com uso de árvores de decisão.

Autor/Ano	Variável predita	Variável preditora	Técnica preditiva	Área (km²)	País/Região
Behrens et al. (2010)	Classes de solo	Relevo	ACL	300	Alemanha
Bou-Kheir et al. (2010)	COS	Relevo, Geologia	ACL	5.748	Dinamarca
Brown (2007)	Argilominerais, COS	Relevo, Rad. espec. (SR)	ADr	Várias áreas	Mundo
Bui &Moran (2003)	Classes de solo, Geologia	Relevo, Rad. espec. (SR)	ACL	$1,1 \times 10^6$	Austrália
Crivelenti et al.(2009)	UM	Relevo, Geologia	AD, ACL	772	Brasil
Giasson et al. (2013)	UM	Relevo	AD, ACL	532	Brasil
Grinand et al. (2008)	UM	Organismos, Relevo, Geologia	ACL, ADr	900	Europa
Lacoste et al. (2011)	Classes de solo, Geologia	Relevo, Geologia, Rad. espec. (SR)	ACL (ADr)	27.020	França
Lagacherie & Holmes (1997)	UM	Relevo, Geologia	ACL	35	França
Lemercier et al. (2012)	Geologia, Classes de solos por drenagem	Relevo, Geologia, Rad. espec. (AF)	ACL (ADr)	4.645	França
Minasny & McBratney (2007)	Classes de solo	Relevo, Rad. espec.(AF e SR)	AD, ACL	Várias áreas	Austrália
Moran e Bui (2002)	Classes de solo, Geologia	Relevo, Rad. espec. (SR)	ACL (ADr)		Austrália
Vasques et al. (2008)	COS	Rad. espec.(SR)	ADr, AR	3.585	EUA

COS: carbono orgânico do solo; UM: unidades de mapeamento de solo; Rad. espec.: Radiação espectral; SR: sensores remotos; AF: aerofotogrametria; AD: árvore de decisão; ACL: árvore de classificação; ADr: árvore de decisão reforçada; AR: árvore de regressão.

Os valores da acurácia geral de trabalhos de predição de solos com uso de árvores de decisão variam conforme a complexidade da região de estudo, da quantidade de classes a serem preditas, das variáveis preditoras, do tamanho do conjunto de dados e do algoritmo de árvore de decisão empregado, mas de uma forma geral, os resultados das pesquisas da Tabela 3 têm mostrado uma porcentagem de classificações corretas em torno de 60,0%. Para aumentar o poder preditivo da árvore de decisão, alguns autores têm

aderido ao uso da árvore de decisão reforçada ou “*decision tree boosting*”. Este processo de reforço é baseado em Friedman (2001), o qual relata melhorias na acurácia de modelos de árvore de classificação. Segundo Adhikari et al. (2014), este processo consiste em gerar muitas árvores de decisão a partir do mesmo conjunto de dados, calculando pesos para cada árvore (com base em sua acurácia) e as combinar em uma única predição. Moran & Bui (2002) aplicaram este procedimento em árvore de classificação para mapear os tipos de solo e relataram que o erro de classificação foi minimizado pelo uso deste método. Outros exemplos de uso desta árvore de classificação incluem os estudos de Lacoste et al. (2011) e Lemercier et al. (2012), que avaliaram a predição de ocorrência de tipos de solo e material de origem na França, os quais sugeriram que a combinação de árvores com 10 classificadores reduziu a taxa de erro em cerca de 25%.

2.3 Considerações Finais

As técnicas preditivas vêm sendo utilizadas em diversos setores privados e governamentais, porém mapeamentos de solos com estas técnicas tem seu primeiro registro no Brasil somente no ano de 2006 com o trabalho de Giasson et al. (2006), e a utilização das árvores decisão no ano de 2009 sob autoria de Crivelenti et al. (2009). Enquanto que as pesquisas internacionais com MDS e árvores de decisão são datadas desde as décadas de 80, no Brasil isto é ainda muito incipiente e são poucos os profissionais que detêm conhecimento para sua utilização (ten Caten et al., 2013; Omuto et al., 2013). Por isso, ainda são poucos os trabalhos publicados nessa linha de pesquisa, bem como literaturas de fundamentos, exemplificando o emprego dos algoritmos.

No MDS, diversos algoritmos de árvores de decisão têm sido testados, porém ainda não existe um protocolo de procedimentos definindo quais algoritmos devem ser utilizados devido às adaptações e melhorias frequentemente realizadas nestes algoritmos e também ao surgimento de novos algoritmos de árvores de decisão (Friedman, 2001; Basgalupp, 2010). Além das vantagens das árvores de decisão supracitadas, no Brasil a predição de ocorrência de solos têm predominado o uso dos algoritmos de árvore de

decisão C4.5 ou J48 (Quinlan, 1993) e CART (Breiman et al., 1984). Assim, esta revisão apresentou procedimentos matemáticos utilizados pelos algoritmos C4.5 (J48) e CART, com exemplos de subdivisões dos dados usando o índice Gini. Todavia, alguns problemas devem ser superados em qualquer que seja o algoritmo de decisão/classificação utilizado, tais como as estratégias de limitar o crescimento da árvore, diminuir o erro resultante da classificação, gerar árvores de menor tamanho com maior acurácia, integrar o conhecimento prévio. Portanto, embora as árvores de decisão apresentem diversas vantagens para uso no mapeamento de solos, mais estudos se tornam necessários para diminuir estas limitações preditivas e melhorar os resultados dos modelos de predição de ocorrência de solos gerados por árvores de decisão.

3. CAPÍTULO II - COMPARAÇÃO DO USO DE MODELOS DIGITAIS DE ELEVAÇÃO EM MAPEAMENTO DIGITAL DE SOLOS EM DOIS IRMÃOS, RS, BRASIL

3.1 Introdução

Um modelo digital de elevação (MDE) é definido como a representação quantitativa digital da variação contínua do relevo sobre o espaço (Moore et. al., 1993). Como a relação entre a ocorrência dos solos e os atributos da paisagem é um conceito consolidado na ciência do solo, os MDE são fundamentais para o mapeamento digital de solos (MDS), pois são fontes digitais disponíveis adequadas para correlacionar a ocorrência e distribuição de solos com as feições do terreno (Florinsky, 2012). Devido à fundamental influência que o relevo exerce na formação dos solos e da ampla disponibilidade dos MDE (Behrens et al., 2010), os atributos do terreno derivados dos MDE são comumente utilizados como variáveis preditoras ao estabelecer as relações solo-paisagem pelas técnicas do MDS (Lagacherie & McBratney, 2007). Desta forma, o desenvolvimento do MDS está diretamente relacionado com os avanços tecnológicos na captação, geração e disponibilidade dos MDE (Cavazzi et al., 2013).

A partir de um MDE podem ser derivados os atributos do terreno primários e secundários. Os primários são gerados diretamente a partir do MDE, tais como elevação, declividade e orientação das vertentes, entre outros.

Os secundários são resultantes de análises geomorfométricas mais elaboradas da paisagem como é o caso do índice de umidade topográfica, índice de convergência topográfica e o índice de posição topográfica (Wilson & Gallant, 2000; Tagil & Jenness, 2008). Contudo, os MDE adquiridos por sensores remotos orbitais apresentam erros e têm sido constantemente melhorados e/ou refinados com uso de algoritmos que filtram os picos anômalos, eliminação de falsas depressões e pontos ausentes de informação e melhorias da definição de corpos d´água e linhas de encosta (Rodrigues et al., 2010; Wilson, 2012). Os projetos Brasil em Relevo (Miranda, 2005) e o TOPODATA (Valeriano & Rossetti, 2012) visam aprimorar os dados originais do SRTM para o Brasil, corrigindo as falhas e distorções ou refinando o tamanho do pixel. A nível mundial, é disponibilizado o MDE SRTM v4.1 (*Shuttle Radar Topographic Mission* v4.1) com preenchimento dos pontos vazios originais do SRTM (Jarvis et al., 2008) e a segunda versão do projeto *Advanced Spaceborne Thermal Emissionand Reflection Radiometer* (ASTER GDEM v2) (Meyer et al., 2012).

Segundo Cavazzi et al. (2013), o meio ambiente não pode ser estudado, modelado ou visualizado em toda a sua complexidade e detalhes e, por isso, se utiliza do efeito de escala para selecionar e generalizar as informações. Segundo McBratney et al. (2003), em se tratando de mapas digitais de solos, diferentemente da cartografia convencional, a escala é um conceito complexo sendo melhor substituído por resolução espacial e tamanho de pixel. Devido à disponibilidade de diferentes tipos de MDE pode acreditar-se que os MDE com maior resolução espacial (menor tamanho de pixel) possibilitam uma melhor representação das características da paisagem e, com isso, gerar modelos preditores de ocorrência de tipos de solos mais acurados. McBratney et al. (2003) apresentam aproximações compatíveis entre o espaçamento de pixel e escalas de mapas. Todavia, além destas relações cartográficas, Smith et al. (2006) relatam que a acurácia dos trabalhos com MDS pode ser influenciada pelas características da paisagem e pela forma como estas informações são representadas pelos MDE.

Os estudos que abordam a influência dos tipos e resoluções dos MDE sobre modelos preditivos de solos são mais comumente encontrados quando envolvem a predição espacial de propriedades dos solos, do que os estudos de predição de ocorrência de solo (Behrens et al., 2010). Em trabalho

recente, Giasson et al. (2013) relatam que a comparação entre três MDE, com pixel de 30 m e 90 m, utilizados para a predição de unidades de mapeamento de solos (UM) não foi conclusiva, pois a diferença na capacidade de predição dos modelos foi muito pequena. Ao utilizarem a análise wavelet e variáveis preditoras derivadas do MDE TOPODATA, ten Caten et al. (2012) sugerem que a resolução espacial a ser utilizada no MDS deva ser entre 32 e 40 m. Em estudo realizado por Cavazzi et al. (2013), os autores demonstraram que em áreas com relevo menos declivoso, os MDE com menores resoluções espaciais (140 m) geraram modelos preditores de solos mais acurados.

Neste contexto, foram avaliados e comparados modelos de predição de ocorrência de solos gerados por árvore de decisão, correlacionando as UM extraídas de um mapa convencional de solos na escala de 1:20.000 e 12 atributos do terreno derivados de seis tipos e resoluções de MDE.

3.2 Material e Métodos

A área de estudo foi o município de Dois Irmãos, situado na região fisiográfica da Encosta Inferior do Nordeste do Estado do Rio Grande do Sul. Esta região é caracterizada por compor parte da borda do relevo do Planalto das Araucárias, cuja geologia é representada pelas rochas basálticas da Formação Serra Geral e pelos arenitos da Formação Botucatu (GEOBANK, 2014). O município ocupa uma área aproximada de 68,5 km^2 e apresenta quatro classes principais de relevo: plano, suave ondulado, ondulado e forte ondulado, mas predominam as formas de relevo menos íngremes (plano e suave ondulado). O município situa-se no limite de dois tipos climáticos, segundo a classificação do clima de Köppen: Cfa (clima subtropical úmido quente) e Cfb (clima subtropical úmido temperado). O relevo foi considerado como o principal fator diferenciador entre os tipos de solo, pois além de condicionar os fluxos de água, a drenagem, o acúmulo de materiais e os processos erosivos, as informações espaciais do relevo estão disponibilizadas em maior quantidade e de livre acesso, tais como as variáveis geradas a partir dos MDE. Assim, foram utilizados somente atributos do terreno como variáveis preditoras de ocorrência de solos, corroborando com o objetivo deste estudo que foi avaliar a influência dos MDE sobre modelos preditores de ocorrência de

solos gerados por árvore de decisão.

Os MDE utilizados para derivar os atributos do terreno estão representados na Figura 6 e são: a) ASTER GDEM v2, com resolução espacial de 30 m, apresenta um acréscimo de 260.000 pares estereoscópicos de imagens em relação à primeira versão do ASTER GDEM (Meyer et al., 2012); b) SRTM v4.1, disponibilizado para a América do Sul com tamanho de pixel de aproximadamente de 90 m (Jarvis et al., 2008); c) TOPODATA, em que os dados originais do SRTM foram interpolados para refinamento no tamanho do pixel de 90 para 30 m, além de inclusão de informação nos pontos vazios (Valeriano & Rossetti, 2012); d) Brasil em Relevo (BR_Relevo), com o mesmo tamanho de pixel que o dado original do SRTM (90 m), mas com correções de cortes e preenchimentos (Miranda, 2005) e; e) os MDE gerados nas resoluções de 30 m (CN30) e 90 m (CN90), com auxílio da função "*topo to raste*r" do programa ArcGis 9.3 (ESRI, 2009) a partir das curvas de nível com equidistância vertical de 20 m disponibilizadas em Hasenack & Weber (2010). Para comparar e avaliar qualitativamente os MDE foram gerados dois perfis de elevação nos sentidos Norte-Sul (A-A') e Leste-Oeste (B-B') (Figura 6), a partir de dois transectos traçados em ambiente de SIG.

A correlação entre as características da paisagem e a distribuição das UM no município de Dois Irmãos foi realizada pelo treinamento do algoritmo de árvore de decisão SimpleCart, a partir de informações coletadas em ambiente de sistema de informação geográfica (SIG) da variável resposta (UM) e das variáveis preditoras (atributos do terreno). As UM foram extraídas do mapa de solos na escala 1:20.000 (Figura 7), que faz parte do Levantamento de Reconhecimento com Detalhes dos Solos do Município de Dois Irmãos (Klamt et al., 1993).

A partir de cada MDE foram gerados doze atributos do terreno, sendo nove primários (elevação, declividade, curvatura, curvatura em perfil, curvatura planar, direção do fluxo, comprimento do fluxo, acúmulo do fluxo, orientação das vertentes) e três secundários (índice de umidade topográfica, índice de convergência topográfica e índice de posição topográfica). A variável distância horizontal das redes hidrográficas foi calculada a partir da rede hidrográfica da base cartográfica vetorial 1:50.000 (Hasenack & Weber, 2010). Para gerar os modelos de árvores de decisão (AD) foram coletados os dados

dos atributos do terreno e das UM em 4.280 pontos amostrais distribuídos aleatoriamente por toda a área de estudo. Esta amostragem corresponde a um ponto para cada área mínima mapeável para a escala 1:20.000 (1 ponto/1,6 ha). Os dados foram tabelados e exportados para o programa de mineração de dados Weka 3.6.3 (Hall et al., 2009), a fim de serem treinados pelo algoritmo SimpleCart com número de elementos no nó final (M) igual a 2 (M2) e 11 (M11). Foram testados diversos valores para M (dados não mostrados), sendo que M igual a 11 (M11) gerou AD menores e menos complexas sem reduzir demasiadamente a acurácia dos modelos e a quantidade das UM preditas.

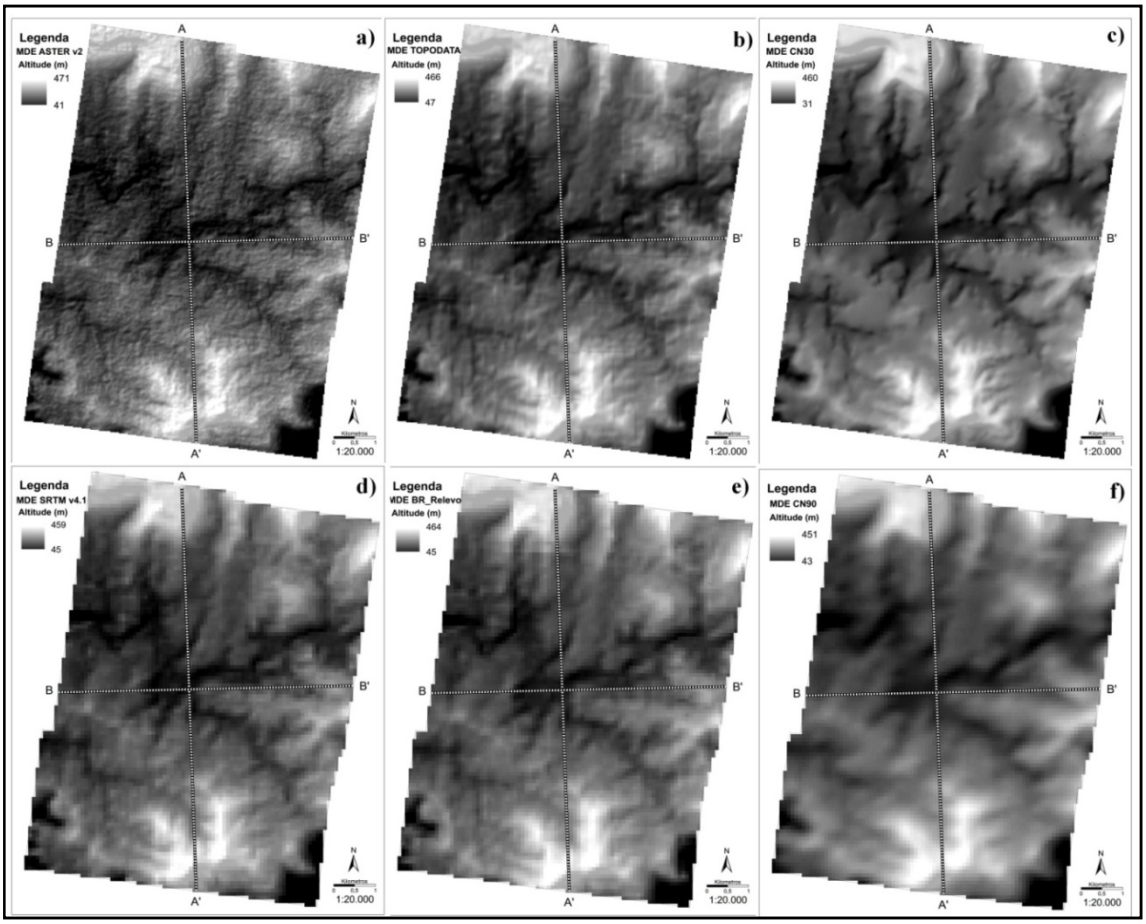

Figura 6. Transectos e modelos digitais de elevação utilizados: a) ASTER GDEM v2; b) TOPODATA; c) MDE CN30; d) SRTM v4.1; e) BR_Relevo e; f) CN90.

A validação dos modelos preditores foi realizada com a totalidade dos dados da área de estudo. Para isso, foram coletadas todas as informações das variáveis preditoras (atributos do terreno) e das variáveis reposta (UM) ocorrentes em todos os pixeis. Estes dados foram tabulados e importados para

o programa Weka 3.6.3 (Hall et al, 2009) para aplicação do método de validação de modelos com dados independentes (*Supplied test set*). O método de validação com dados independentes é realizado por comparações em matriz de erros (Congalton, 1991), confrontando a predição de ocorrência de solos com a distribuição original (Rossiter, 2004). Assim, foram obtidas as estimativas da acurácia geral (AG), que é a proporção de instâncias corretamente classificadas, e o índice Kappa (Cohen, 1960) que mede as concordâncias compensando o acaso.

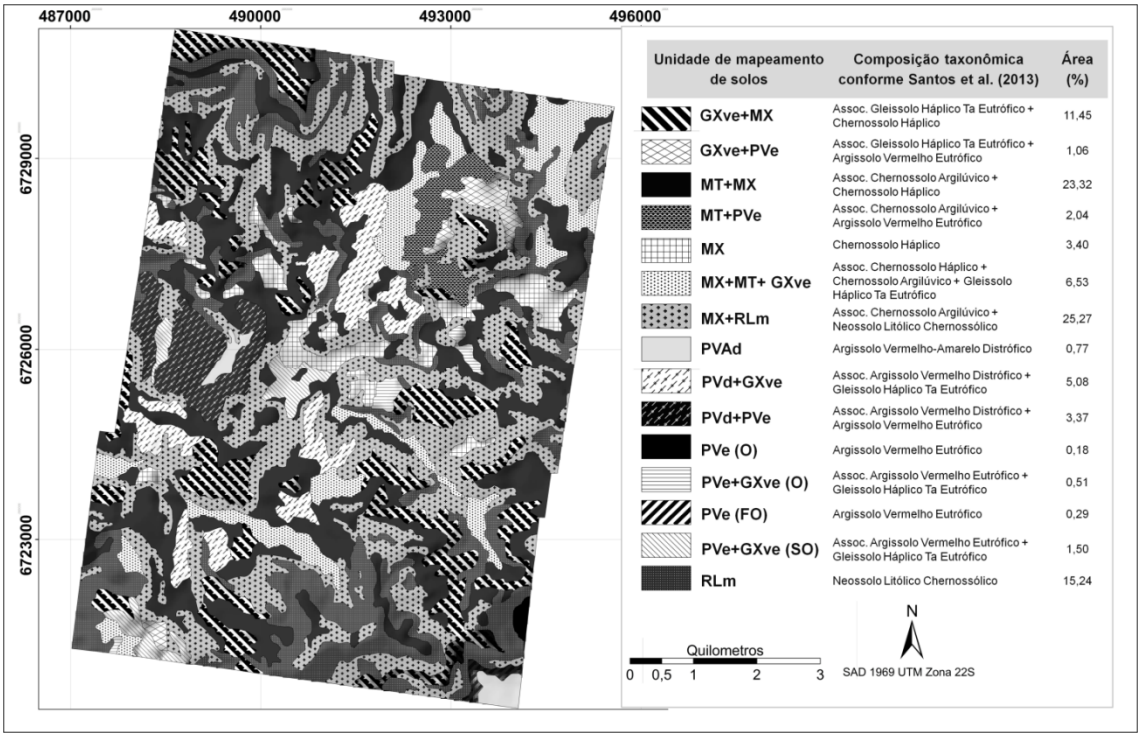

Figura 7. Mapa de solos do município de Dois Irmãos, Rio Grande do Sul (Klamt et al., 1993).

3.3 Resultados e Discussão

A representação do relevo da área de estudo foi diferenciada para cada MDE utilizado, os quais apresentaram diferentes valores mínimos e máximos para elevação (Figura 6). Os maiores valores de elevação foram encontrados nos MDE ASTER GDEM v2 e o TOPODATA e os menores valores de elevação nos MDE CN30 e ASTER GDEM v2. As maiores amplitudes nos valores da elevação foram encontradas nos MDE com resolução espacial de 30

m, principalmente, no MDE ASTER GDEM v2. Segundo Guth (2010), dentre os MDE gerados a partir de dados obtidos por sensores remotos orbitais, o ASTER GDEM apresenta maior quantidade de artefatos e variações superficiais que são consideradas e quantificadas como valores de elevação (Rodrigues et al, 2010; Wilson, 2012).

Os perfis de elevação dos transectos (Figuras 8 e 9) demonstram que os MDE gerados a partir da carta topográfica na escala de 1:50.000 (CN30 e CN90) apresentaram os menores valores de elevação em praticamente toda extensão destes transectos. Estes resultados estão em concordância com o trabalho de Chagas et al. (2010), que ao compararem diferentes MDE com resolução espacial de 30 m constataram que o MDE gerado a partir de curvas de nível, com equidistância vertical de 20 m na escala de 1:50:000, permitiu representar melhor as características do terreno e apresentou menores valores de elevação do que os MDE ASTER GDEM e SRTM.

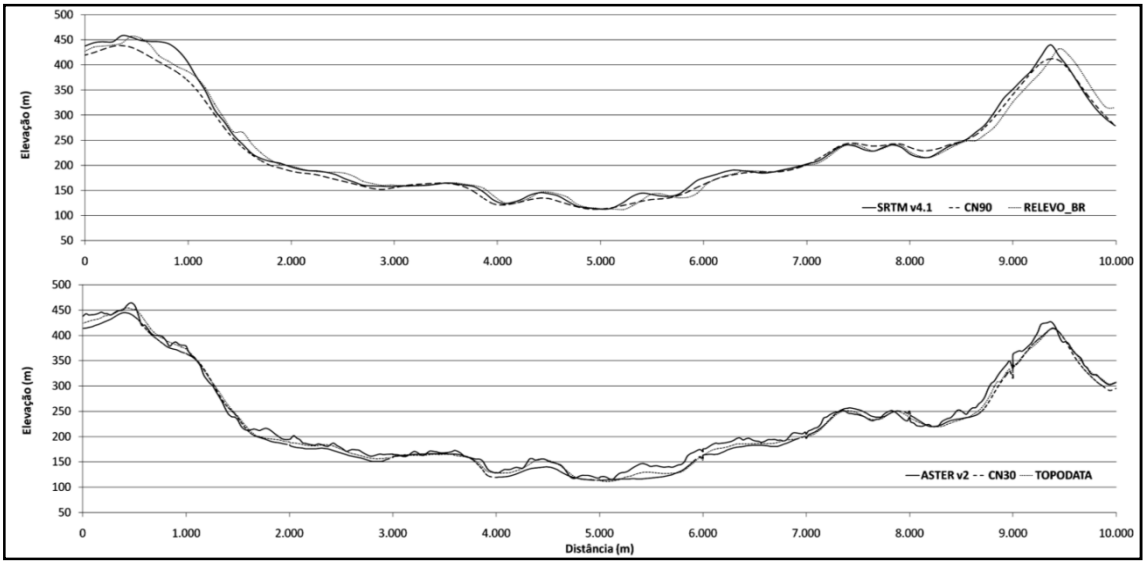

Figura 8. Perfis de elevação no transecto A–A` dos modelos digitais de elevação com resolução espacial de 90 m (SRTM v4.1, CN90 e BR_Relevo) e com resolução espacial de 30 m (ASTER v2, CN30 e TOPODATA).

Os MDE originados de dados orbitais do SRTM v4.1 e do ASTER GDEM v2 apresentaram os maiores valores de elevação e as maiores oscilações nestes valores, indicando assim, a influência das variações superficiais causadas por edificações e árvores. Embora alguns trabalhos

indiquem que o MDE ASTER GDEM apresenta algumas vantagens em relação aos produtos do SRTM, tal como a resolução espacial de 30 m (Rodrigues et al., 2010), as maiores oscilações nos valores de elevação do MDE ASTER GDEM v2 indicou que este MDE apresenta uma maior quantidade de artefatos do que o SRTM, sendo agravados em regiões de relevo mais suave resultando numa pior representação do relevo, tal como constatado por Guth (2010).

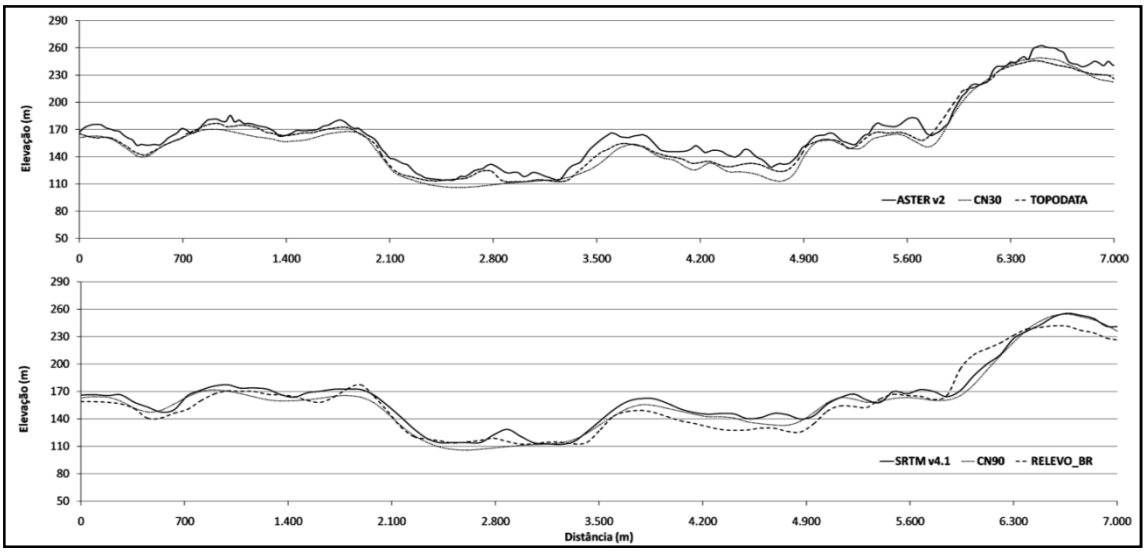

Figura 9. Perfis de elevação no transecto B–B` dos modelos digitais de elevação com resolução espacial de 90 m (SRTM v4.1, CN90 e BR_Relevo) e com resolução espacial de 30 m (ASTER v2, CN30 e TOPODATA).

Os resultados da avaliação da acurácia dos modelos indicaram que as predições de ocorrência de UM foram influenciadas tanto pelo tipo como pela resolução espacial dos MDE (Tabela 4). Os modelos preditores gerados a partir dos MDE com pixel de 90 m, utilizando M2, foram os mais acurados. Estes modelos apresentaram as maiores AD, o que permitiu estabelecer melhores correlações entre as características do terreno e as UM, resultando na predição de todas as 15 UM presentes no mapa convencional de solos. A correlação das UM com os atributos do terreno derivados do MDE CN90 resultou um modelo preditivo com AG de 54,71 % e índice Kappa de 0,46. O modelo preditor de UM gerado com os dados do MDE SRTM v4.1 apresentou uma AG de 52,69 % e índice Kappa igual a 0,43. Estes resultados da acurácia são comparáveis e consistentes com demais estudos com MDS, tal como os resultados obtidos por Cavazzi et al. (2013), em que os mapas preditores de

classes de solo apresentaram acurácia entre 35 % e 60 % ao utilizar MDE gerados a partir de curvas de nível, na escala de 1:50:000. Behrens et al. (2010) analisaram digitalmente um terreno em multiescala, utilizando filtros com dimensões que variaram de 3 × 3 até 31 × 31 pixeis e encontraram uma acurácia geral de 54 % na área de validação e, também com os resultados obtidos por Giasson et al. (2013), que ao compararem diferentes algoritmos de treinamento e MDE encontraram valores de acurácia entre 50,0 % e 57,3 %.

Tabela 4. Resultados da avaliação da acurácia dos modelos de predição de ocorrência de unidades de mapeamento de solos usando seis modelos digitais de elevação.

MDE*	Res. espacial	M2				M11			
		AG (%)	Kappa	Tam.	Nº UM	AG (%)	Kappa	Tam	Nº UM
ASTER	**30 m**	37,16	0,23	141	11	36,66	0,22	73	9
TOPODATA	**30 m**	45,98	0,35	459	13	44,28	0,32	117	9
CN30	**30 m**	45,65	0,34	431	14	43,61	0,31	169	12
BR_Relevo	**90 m**	47,93	0,37	921	15	44,48	0,33	195	12
SRTM v4.1	**90 m**	52,69	0,43	999	15	47,06	0,36	107	11
CN90	**90 m**	54,71	0,46	985	15	46,61	0,35	147	13

*MDE = modelo digital de elevação; M = número de elementos no nó final da árvore de decisão; AG = acurácia geral; Tam.= tamanho da árvore de decisão; Nº UM: quantidade de unidades de mapeamento de solos estimada em cada modelo preditor.

Nos modelos preditores gerados a partir dos MDE com tamanho de pixel de 30 m e M2, a quantidade de UM preditas variou conforme o tipo de captação dos dados de elevação. Ao utilizar os atributos do terreno derivados do MDE CN30 o modelo preditor estimou a ocorrência de 14 UM, com o MDE TOPODATA foram preditas 13 UM e com uso dos atributos derivados do MDE ASTER GDEM v2 foram estimadas 11 UM. O tamanho das AD geradas com os MDE de 30 m foi consideravelmente menor do que as AD geradas com os MDE de 90 m e M2, principalmente, para o MDE ASTER GDEM v2 que apresentou as maiores oscilações nos valores da elevação.

Os modelos preditores de UM gerados com M11 apresentaram AD menores, foram menos acurados e estimaram menos UM do que os gerados com M2. Estas diferenças foram mais expressivas nos modelos mais acurados (CN90 e SRTM v4.1), indicando que a redução no tamanho das AD pelo aumento do valor de M pode prejudicar a capacidade preditiva de modelos, principalmente, os mais acurados. Todavia, os modelos gerados a partir dos

MDE com pixel de 90 m (CN90 e SRTM v4.1) permaneceram sendo os mais acurados, enquanto o modelo preditor de UM gerado a partir do MDE ASTER GDEM v2 permaneceu sendo o menos acurado e com a menor quantidade de UM preditas. Assim, indiferentemente do valor de M adotado, os modelos de predição de ocorrência de UM gerados a partir dos MDE com tamanho de pixel de 90 m estabeleceram as melhores correlações solo-paisagem e, com isso, geraram os modelos de predição de ocorrência de UM mais acurados.

Os atributos do terreno derivados dos MDE elevação, declividade, comprimento de fluxo e orientação das vertentes foram os que melhor explicaram a relação solo-paisagem em todos os modelos preditores. A importância destas variáveis é constatada por diversos autores (Behrens et al., 2010; ten Caten et al., 2012; Giasson et al., 2013), que comumente têm identificado a elevação e a declividade dentre os principais atributos do terreno usados no MDS. Utilizando os MDE ASTER GDEM, TOPODATA e SRTM, ten Caten et al. (2012) e Giasson et al. (2013) constataram que, além da elevação e da declividade, a curvatura, a curvatura em perfil, a direção de fluxo, o acúmulo de fluxo e o índice de umidade topográfica também apresentaram contribuição expressiva para explicar a ocorrência de classes de solo. Isto está relacionado ao fato da elevação apresentar um importante papel na definição do clima local e os atributos declividade, orientação das vertentes e comprimento de fluxo por influenciar a velocidade do fluxo superficial e subsuperficial de água, evapotranspiração, insolação, teor de água no solo, processos erosivos e deposicionais, afetando assim, os processos pedogenéticos, propriedades e o potencial agrícola dos solos (Moore et al., 1993; Wilson & Gallant, 2000).

Ao analisar os dados utilizados para treinamento em gráficos de caixa - *box plot* - (Figura 10), pode-se observar que as amostragens apresentaram dados da elevação muito semelhantes entre os MDE de 30 m (representados pelo MDE ASTER GDEM v2) e nos MDE de 90 m (representados pelo MDE CN90). Os dados da elevação, analisados pela amplitude, média e mediana, foram diferenciados entre as UM, demonstrando a alta relação da distribuição das UM com as cotas altimétricas. Todavia, alguns delineamentos como a UM MT+MX, a UM GXve+MX, a UM RLm e a UM MX+RLm ocorrem em variadas elevações, apresentando maiores amplitudes e

mais valores extremos (*outliers*), sendo um dos motivos para seleção dos demais atributos do terreno (como a declividade e o comprimento de fluxo) na construção das AD e geração dos modelos preditores.

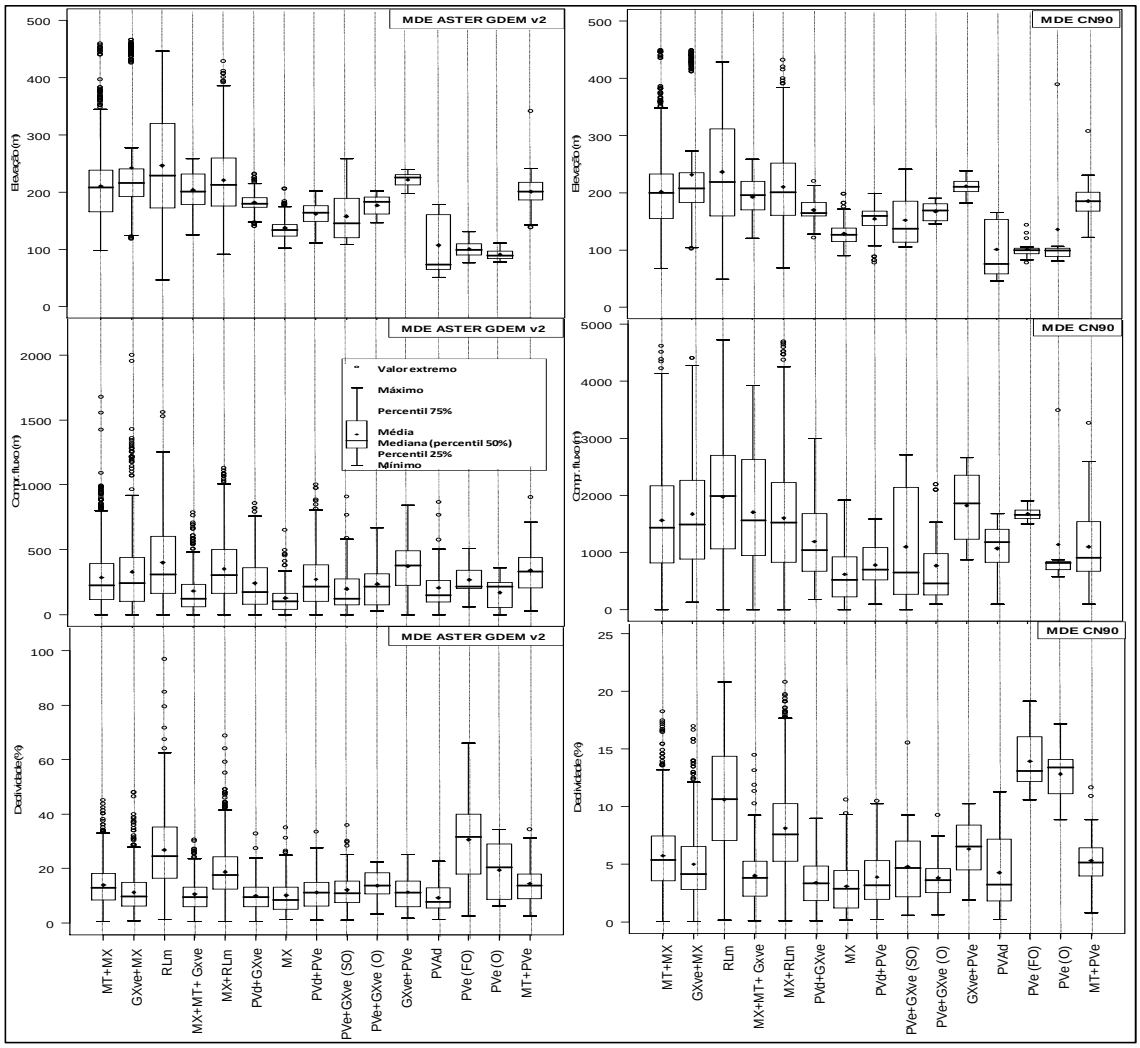

Figura 10. Gráficos de caixas dos dados de elevação, comprimento de fluxo e declividade derivados dos modelos digitais de elevação ASTER GDEM v2 (30 m) e CN90 (90 m) para cada unidade de mapeamento constante no mapa convencional de solos.

A declividade é considerada como um atributo local sendo calculada com base nos valores da elevação e no tamanho dos seus pixeis vizinhos (3 x 3 pixeis) (Florinsky, 2012). Ao utilizar os MDE com maior tamanho de pixeis, os valores da declividade foram menores para os mesmos valores de elevação. Isto pode ser observado nos gráficos de caixas, em que o MDE ASTER GDEM v2 apresentou os maiores valores de declividade (95 %), enquanto que para o MDE CN90 o valor máximo para a declividade foi de 23 %. Em relação ao

comprimento de fluxo, este é considerado como um atributo não-local e é uma medida da distância ao longo do trajeto de fluxo (determinada pela direção do fluxo) a partir de uma dada célula até a sua saída da bacia de drenagem (Florinsky, 2012).

Os valores para o comprimento de fluxo, assim como a declividade, foram diferentes entre os MDE de 30 m e de 90 m. No MDE ASTER GDEM, os valores deste atributo foram menores, pois percorreram menores distâncias (menor tamanho de pixel) e se apresentaram mais assimétricos e com maior quantidade de valores extremos dentro de cada UM do que no MDE CN90, que por sua vez, apresentou maiores diferenças nos valores de média e mediana entre as UM. Desta forma, os MDE com resolução espacial de 90 m possibilitaram gerar modelos mais acurados, pois apresentaram menores quantidades de valores assimétricos e extremos em cada UM, bem como uma maior diferenciação dos valores dos atributos (médios e da mediana) entre as UM, permitindo assim, o modelo preditor realizar separações mais homogêneas dos dados (Rokach & Maimon, 2008) para cada UM e gerar correlações solo-paisagem mais acuradas.

Estes resultados estão em conformidade com os estudos de Thompson et al. (2001) e Smith et al. (2006) em que à medida que a resolução espacial do MDE é reduzida (aumenta o tamanho do pixel), a declividade é diminuída, o comprimento do percurso de escoamento é aumentado e os detalhamentos nos atributos do terreno são perdidos. Contudo, os MDE com menor tamanho de pixel não são necessariamente considerados as melhores fontes de informações para se gerar variáveis preditoras de solos (Cavazzi et al., 2013). Isto decorre por apresentarem uma maior quantidade de dados espacializados e uma variação acentuada nos valores dos atributos do terreno dentro de cada UM. Em se tratando de uma área com predomínio de relevo mais suave, isto resultou em amplitudes dos dados mais semelhantes entre as UM do que às encontradas com os atributos gerados a partir dos MDE com pixel de 90 m. Assim, uma maior amplitude dos dados dentro de cada UM, juntamente com uma menor variação nas características da paisagem entre as áreas ocupadas pelas diferentes UM, resultaram modelos preditores menos acurados, pois dificultaram o treinamento do algoritmo e a separação em dados mais homogêneos para cada classe, neste caso as UM. Adicionalmente, os

MDE com tamanho de pixel de 30 m apresentaram mais oscilações na elevação e, por isso, ao predizer a ocorrência de UM para toda a área, os mapas auxiliares das características do relevo apresentam maior quantidade oscilações e de informações divergentes das regras de classificação geradas, diminuindo ainda mais a sua acurácia.

Além da influência das características da paisagem e da forma como estas são representadas pelos MDE, as UM representam áreas ambientalmente semelhantes em uma determinada escala. Assim, ao predizer a ocorrência de UM pelas técnicas do MDS, as UM são também distribuídas para áreas cujas características da paisagem sejam semelhantes conforme os dados de entrada e as correlações realizadas pelo algoritmo. No presente estudo, as informações pedológicas de referência foram extraídas de um mapa convencional de solos na escala de 1:20.000, em que a área mínima mapeável (0,4 cm^2) é de 16.000 m^2 (1,6 ha). Segundo McBratney et al. (2003), o tamanho do pixel sugerido e compatível com mapas na escala de 1:20.000 é de 20 m, variando até 200 m conforme a área e objetivo de estudo. Estes autores ainda relatam que uma resolução espacial mínima deva ser de 2 x 2 pixeis. Assim, os MDE de 90 m representam uma resolução mínima de 32.400 m^2 (180 m x 180 m), enquanto os MDE de 30 m (60 m x 60 m) de somente 3.600 m^2, valor este muito aquém da área mínima mapeável para a escala de 1:20.000. Ao considerarmos que na área de estudo predominam fases de relevo mais suaves e que todas as UM estão representando áreas maiores que a área mínima mapeável, a representação contínua do relevo por MDE com resolução espacial de 30 m não se mostrou a mais indicada, ao ponto destes MDE apresentarem mais oscilações nos valores da elevação (ASTER GDEM), aumentarem demasiadamente a quantidade de informações espacializadas para uma área com relevo menos complexo, o que dificultou o treinamento do algoritmo e a geração de modelos preditores mais acurados. Desta forma, uma maior diferenciação nos valores dos atributos do terreno dentro das áreas de cada UM se torna mais importante do que um maior detalhamento espacializado com uso de MDE com pixel de 30 m, que não possibilitaram gerar modelos de predição de ocorrência de solos mais acurados.

De uma forma geral, os resultados da acurácia dos modelos preditores são satisfatórios e comparáveis com demais estudos de predição

categórica de solos (classes ou unidades de mapeamento de solos). Assim, as correlações solo-paisagem geradas por AD (algoritmo SimpleCart) para predizer a ocorrência de UM para o município de Dois Irmãos, em que predominam fases de relevo plana e suave ondulada, foram melhores quando utilizados atributos do terreno derivados dos MDE com resolução espacial de 90 m, principalmente, com o MDE do projeto SRTM v4.1 e o MDE gerado a partir de curvas de nível (CN90).

3.4 Conclusões

Com o uso de árvores de decisão, os atributos derivados dos MDE que melhor explicam a distribuição espacial das UM para uma área com predomínio de fases do relevo menos complexo foram a elevação, a declividade, o comprimento de fluxo e a orientação das vertentes.

Os MDE com maior resolução espacial oriundos de sensores remotos orbitais apresentam maiores oscilações nos valores da elevação e geram correlações solo-paisagem menos acuradas e menor quantidade de unidades de mapeamento preditas.

Em áreas que predominam relevo plano a suave ondulado, os MDE com resolução espacial de 90 m (SRTM v4.1 e CN90) permitem gerar modelos preditores de ocorrência de UM mais acurados e com maior número de UM preditas.

4. CAPÍTULO III–COMPARAÇÃO DE ESQUEMAS DE AMOSTRAGEM PARA TREINAMENTO DE MODELOS PREDITORES NO MAPEAMENTO DIGITAL DE CLASSES DE SOLOS

4.1 Introdução

O mapeamento digital de solo (MDS) visa correlacionar os solos com os fatores de formação de solos de forma mais quantitativa que os levantamentos convencionais e utilizando modelos quantitativos para inferir as variações espaciais dos solos (Lagacherie & McBratney, 2007). Para isso, os modelos precisam ser treinados e validados com dados que capturem ao máximo a variação espacial dos atributos do terreno e dos solos é necessário o uso de estratégias de amostragem estatisticamente robustas para diminuir os erros da predição (Minasny & McBratney, 2007; Brungard & Boetinger, 2010).

Segundo Brus & Gruijter, (1997), quando realizada a predição de ocorrência de classes ou propriedades de solo por correlação ambiental entre mapas auxiliares (variáveis preditoras) e mapas de referência de solo, a amostragem deve ser fundamentada na teoria da probabilidade. Assim, a amostragem aleatória tem sido comumente utilizada no MDS para amostrar os dados para treinamento por eliminar a subjetividade e permitir a reprodutibilidade simples (Hengl et al., 2003). Todavia, alguns autores têm relatado que classes e unidades de mapeamento de solos (UM) pouco extensas e pouco representativas de uma área não são preditas pelos modelos

quando utilizada a amostragem aleatória (Giasson et al., 2011; ten Caten et al., 2012). Em trabalho realizado por Grinand et al. (2008), os autores utilizaram a amostragem proporcional à área de cada UM, tal como recomendado por Moran & Bui (2002), e desta forma, as UM menos representativas não foram subamostradas, como pode ocorrer na amostragem aleatória simples. A amostragem estratificada também tem sido recomendada para amostrar as UM pouco representativas (Hengl et al., 2003; Stehman, 2008).

Os modelos de predição mais acurados possibilitam gerar mapas digitais de solo mais acurados e que à semelhança dos mapeamentos convencionais de solo os mapas gerados também apresentam erros que devem ser identificados e quantificados (McBratney et al., 2003; Brus et al., 2011). Para isto, Rossiter (2004) aborda diferentes métodos para obtenção da acurácia de mapas temáticos, sendo a comparação do mapa gerado com um mapa de referência uma metodologia comumente empregada no MDS para medir a concordância, tal como realizado por Bui & Moran (2003) e Giasson et al. (2011). Além disso, as estimativas da acurácia podem ser obtidas por métodos avaliação de modelos preditores, os quais fornecem ao usuário uma estimativa dos acertos e dos erros da predição ao final da construção do modelo (Chatfield, 1995; Steyerberg, 2009) e em uma etapa anterior à geração dos mapas digitais (Brus et al., 2011).

O uso de dados independentes e distintos daqueles usados para o treinamento é indicado para avaliar a capacidade preditiva de modelos em diversas áreas do conhecimento (Grinand et al., 2008; Steyerberg, 2009; Brus et al., 2011). Todavia, no MDS tem sido encontrado o uso de diferentes métodos de avaliação de modelos preditores (Grunwald, 2009). Dentre os métodos de avaliação de modelos disponíveis no programa Weka (Hall et al., 2009), o qual tem sido comumente utilizado em trabalhos com MDS no Brasil, estão: a validação cruzada e a validação com divisão percentual, que repartem o conjunto de dados inicialmente amostrados para treinamento em dois subconjuntos, sendo um para treinamento e outro para avaliação; o método de validação aparente, que utiliza todo o conjunto de dados usados no treinamento para a avaliação, e o método de validação com dados independentes e distintos daqueles usados para treinamento.

O objetivo deste estudo foi avaliar e comparar os resultados das

predições de ocorrência de classes de solos geradas por árvore de classificação com dados oriundos de três esquemas de amostragem. Adicionalmente, foram utilizados quatro métodos para avaliar a acurácia dos modelos preditores e os resultados foram confrontados com os valores da concordância dos mapas digitais com o mapa convencional de solos.

4.2 Material e Métodos

O estudo foi realizado na microbacia do Rio Santo Cristo, região Noroeste do Estado do Rio Grande do Sul, e possui uma área de 898 km^2. O clima da região é subtropical úmido, tipo Cfa de Köppen e o material de origem da região é basalto da Formação Serra Geral. O mapa de solos utilizado como referência, na escala de 1:50.000, faz parte do levantamento pedológico e análise qualitativa do potencial de uso dos solos para o descarte de dejetos suínos da microbacia do Rio Santo Cristo (Kämpf et al., 2004).

Os atributos do terreno usados para caracterizar a paisagem foram gerados em ambiente de sistemas de informação geográfica (SIG), com o programa ArcGis 9.3 (ESRI, 2009). Foram derivados sete atributos do terreno (elevação, declividade, direção do fluxo, acúmulo do fluxo, comprimento do fluxo, curvatura e índice de umidade topográfica) a partir do modelo digital de elevação (MDE) ASTER-GDEM v2 com resolução espacial de 30 metros (MEYER et al., 2012). A partir do arquivo vetorial de hidrografia da base contínua do Rio Grande do Sul (Hasenack & Weber, 2010) foi gerada a variável distância horizontal das redes hidrográficas. Todos os mapas foram rasterizados com resolução espacial de 30 metros.

As correlações entre os atributos do terreno e a distribuição espacial dos solos foram geradas por árvore de classificação (Breiman et al., 1984) usando o algoritmo SimpleCart, com número de elementos no nó final (M) igual a 2 (M2), no programa de mineração de dados Weka 3.6.3 (Hall et al., 2009). As árvores de classificação foram utilizadas porque são robustas e permitem o uso de variáveis preditoras tanto nominais como numéricas, além de usar os dados em qualquer escala e com valores discrepantes (Witten et al., 2011; Scull et al., 2003). Para proceder ao treinamento dos modelos preditores, as informações dos atributos do terreno e das unidades de mapeamento (UM)

foram amostradas em 45.000 pontos (1 ponto a cada 2 ha) usando três esquemas de amostragem (Figura 11): i) aleatório simples; ii) aleatório proporcional à área ocupada por cada UM, em que os pontos amostral foram distribuídos aleatoriamente sobre o mapa obedecendo o critério da proporcionalidade; iii) aleatório estratificado pelo número de UM, onde a mesma quantidade de pontos amostrais (4.500) foram distribuídos aleatoriamente dentro da área de cada uma das 10 UM.

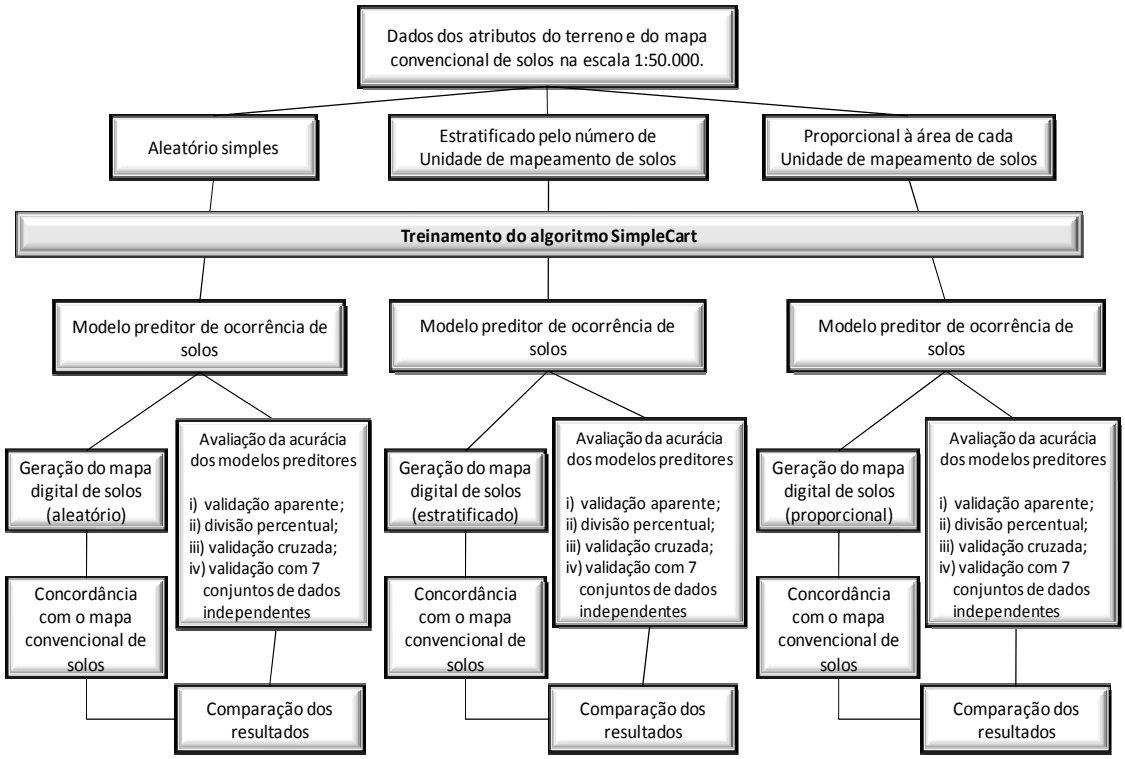

Figura 11. Fluxograma dos métodos usados para geração dos mapas digitais de solos com dados oriundos de três esquemas de amostragem, e a obtenção da acurácia dos modelos preditores e concordância dos mapas gerados.

Os modelos preditores gerados para cada esquema de amostragem tiveram suas acurácias estimadas pelos seguintes métodos de avaliação: i) validação aparente, que se utiliza da totalidade do conjunto de dados usados para treinamento; ii) divisão percentual, sendo 66% dos dados usados para treinamento e 34% para avaliação; iii) validação cruzada com 10 subconjuntos, sendo nove subconjuntos para treinamento e um para avaliação, sendo intercalados até que todos os conjuntos sejam usados para treino e para teste (STEYERBERG, 2009); iv) validação com 7 conjuntos de dados independentes

e distintos dos usados para treinamento. Estes dados independentes foram amostrados aleatoriamente sob diferentes proporções em relação à totalidade da área de estudo (0,75; 1,5; 3,0; 4,5; 6,0; 7,5 e 9,0 %), resultando em conjuntos de dados com 7.500, 15.000, 30.000, 45.000, 60.000, 75.000 e 90.000 pontos amostrais.

As estimativas da acurácia obtidas pelos métodos de avaliação foram comparadas à concordância, obtida pela comparação dos mapas gerados a partir de cada esquema de amostragem com o mapa convencional de solos. Para isto, cada modelo preditor foi transcrito em regras de classificação que foram implementadas em ambiente de SIG para a obtenção dos mapas preditores de ocorrência de solos, sendo um mapa digital para cada esquema de amostragem. Com auxílio da função *Tabulate Area* no ArcGis 9.3, cada mapa digital foi comparado pixel a pixel com o mapa convencional de solos, usando a matriz de confusão de Congalton (1991) para obter os valores da concordância geral, que é a proporção de instâncias corretamente classificadas, e do índice kappa (Cohen, 1960), que mede as concordâncias compensando o acaso.

4.3 Resultados e Discussão

Os esquemas de amostragem resultaram na geração de diferentes modelos preditores de ocorrência de solos com distintos valores da acurácia e da concordância que estão apresentados na Tabela 5. A observação das árvores de classificação geradas permitiu identificar que os atributos do terreno mais importantes para explicar a distribuição espacial dos solos na paisagem foram a elevação, a declividade, o comprimento de fluxo, a distância horizontal das redes hidrográficas e o índice de umidade topográfica, os quais têm sido comumente relatados como importantes variáveis preditoras para o MDS (Behrens et al., 2010; ten Caten et al., 2012; Giasson et al., 2013).

Quando o modelo foi gerado com os dados oriundos da amostragem aleatória, a elevação foi a principal variável preditora de ocorrência de solos, sendo utilizada como nó raiz e em diversos nós internos da árvore de classificação. O comprimento de fluxo, a distância horizontal das redes hidrográficas e a declividade também foram variáveis importantes no modelo

gerado com dados aleatórios, sendo utilizadas como principais nós internos ao separar os dados em subconjuntos mais homogêneos pelo índice Gini do algoritmo SimpleCart. Ao utilizar os dados da amostragem estratificada, a variável elevação também foi usada como a principal variável preditora, enquanto que para os nós internos do modelo foram utilizadas as variáveis declividade, acúmulo de fluxo e o índice de umidade topográfica. Com uso dos dados da amostragem proporcional foi gerada a árvore com menor tamanho, a qual apresentou a variável distância horizontal das redes hidrográficas como nó raiz e para os demais nós internos foram usadas apenas as variáveis elevação e declividade.

Os distintos modelos resultaram diferentes distribuições espaciais dos solos como apresentadas nos mapas digitais da Figura 12. No mapa digital gerado com dados aleatórios (Figura 12b) foram estimadas as seis UM mais extensas e representativas da microbacia do Rio Santo Cristo (G1, LV1, LV2, RR1, RR2 e RR3). Assim como constatado em diversos trabalhos com MDS, a predição de ocorrência de UM pouco representativas é prejudicada quando utilizada a amostragem aleatória (Giasson et al., 2011; ten Caten et al., 2012). Todavia, a distribuição espacial das UM foi a mais semelhante em relação ao mapa convencional, resultando na maior concordância (63,0 %) com o mapa convencional de solos e índice kappa de 0,46.

Os mapas digitais gerados a partir de dados amostrados proporcionalmente à área de cada UM e de forma estratificada resultaram nos menores valores da acurácia. Ao utilizar os dados da amostragem proporcional, que embora garanta a amostragem de todas as UM, resultou num mapa digital (Figura 12d) com as 6 UM mais extensas da microbacia com uma concordância geral de 51,0 % e índice kappa de 0,26. A partir dos dados amostrados de forma estratificada, o mapa preditor de solos estimou todas as 10 UM (Figura 12c), porém com os menores valores de concordância geral (29,2 %) e índice kappa (0,14) devido às UM menos representativas terem sido preditas para áreas maiores que àquelas originalmente mapeadas no mapa convencional. Estes resultados são comparáveis e concordantes com estudos anteriores com MDS, tais como o realizado por Giasson et al. (2011), que encontraram valores de concordância geral de 68,0 % e índice kappa de 0,54 ao compararem o mapa gerado com o mapa convencional de solos, e o de Bui & Moran (2003),

que encontraram valores de acurácia geral entre 53,0 % a 79,0 % e índice kappa 0,33 a 0,74 conforme a escala do mapa convencional e das sub-regiões daquela área de estudo.

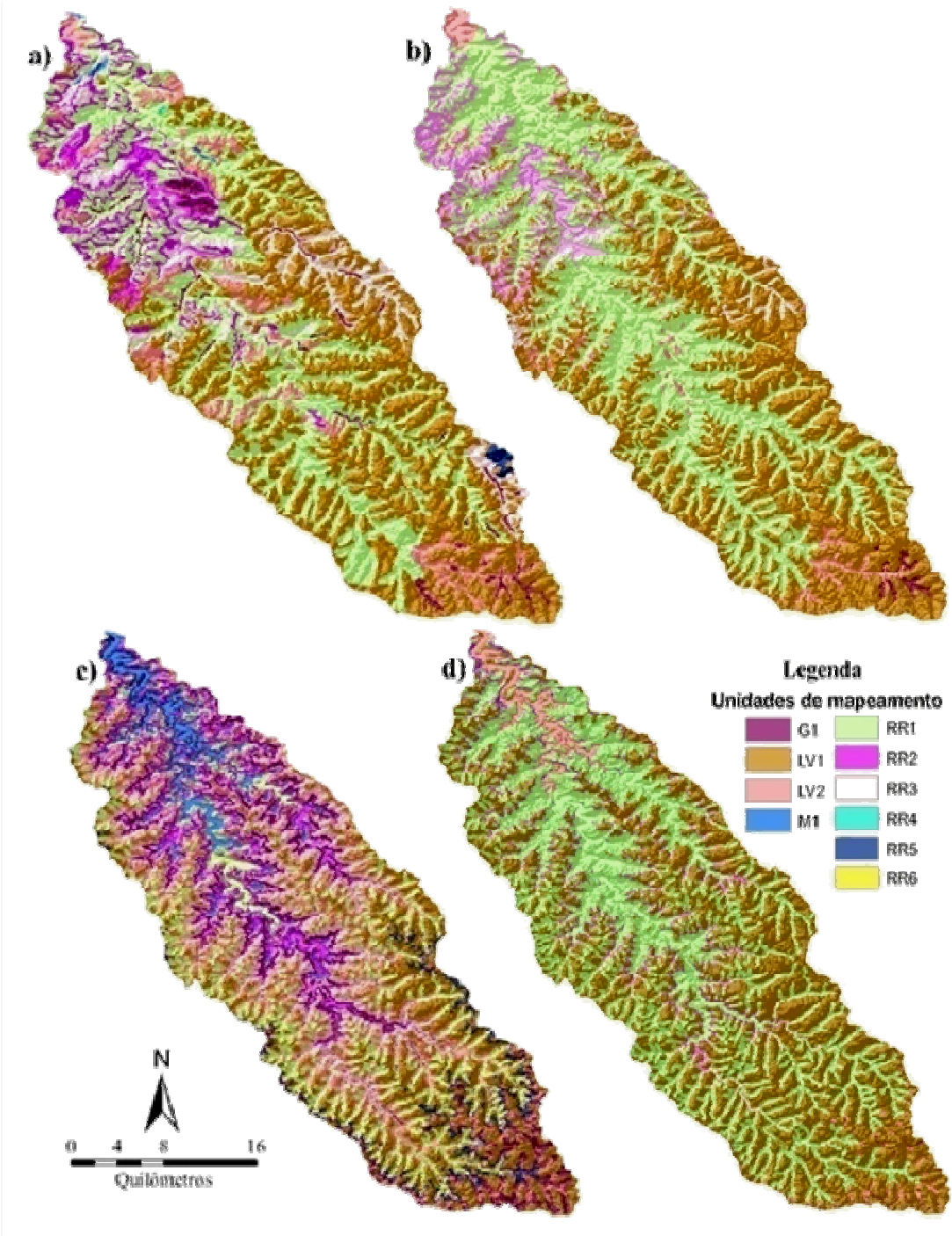

Figura 12. a) mapa convencional de solos da microbacia do Rio Santo Cristo (Kämpf et al., 2004); b) mapa digital de solos gerado com amostragem aleatória simples; c) mapa digital de solos gerado com amostragem aleatória estratificada); d) mapa digital de solos gerado a partir dos dados coletados proporcionalmente à área de cada unidade de mapeamento.

Os resultados da avaliação da capacidade preditiva dos modelos preditores estão dispostos na Tabela 5. Os valores da concordância geral e do índice kappa foram diferenciados para cada modelo gerado pelas combinações de esquema de amostragem e dos métodos de avaliação aplicados. A acurácia dos modelos preditores com uso dos dados independentes e distintos resultou valores muito semelhantes aos encontrados pela concordância, indiferentemente do esquema de amostragem e do tamanho do conjunto de dados independentes. Estes resultados estão em conformidade com a indicação da avaliação com uso de dados independentes e distintos daqueles usados para treinamento dos algoritmos (Grinand et al., 2008; Steyerberg, 2009; Brus et al., 2011).

Tabela 5. Resultados obtidos pelos métodos de avaliação dos modelos e pela concordância entre cada mapa gerado e o mapa convencional de solos.

Método de avaliação		**Esquema de amostragem**					
		Aleatória simples		**Estratificada aleatória**		**Proporcional aleatória**	
		AG[1] (%)	**kappa**	**AG (%)**	**kappa**	**AG (%)**	**kappa**
Avaliação Interna	VC-10	62,1	0,45	66,0	0,62	55,6	0,33
	DP	61,3	0,44	65,9	0,63	55,8	0,33
	VA	64,0	0,46	67,1	0,64	57,0	0,35
	Média	62,5	0,45	66,3	0,63	56,1	0,34
Avaliação com dados independentes	7.500	63,0	0,46	29,6	0,14	51,4	0,27
	15.000	62,3	0,45	29,0	0,13	50,1	0,26
	30.000	62,9	0,46	29,4	0,14	51,7	0,27
	45.000	63,5	0,46	29,0	0,13	51,3	0,26
	60.000	62,6	0,45	29,5	0,14	51,3	0,27
	75.000	62,6	0,45	29,5	0,14	51,3	0,27
	90.000	63,0	0,46	30,1	0,15	51,6	0,27
	Média	62,8	0,46	29,4	0,14	51,2	0,27
Concordância		63,0	0,46	29,2	0,14	51,0	0,26

[1]AG = acurácia geral; VC-10 = validação cruzada com 10 subconjuntos; DP = divisão percentual; VA = validação aparente.

Ao realizar a avaliação com os métodos de validação aparente, cruzada e por divisão percentual, a acurácia dos modelos gerados com dados oriundos da amostragem estratificada e proporcional foi superestimada, enquanto que a avaliação da acurácia do modelo preditor gerado com os dados

da amostragem aleatória resultou valores semelhantes aos obtidos pela concordância e pelo uso de dados independentes. Assim, o uso da amostragem aleatória se mostrou mais robusta e vantajosa para a predição de ocorrência de UM por gerar correlações solo-paisagem mais acuradas, cujos valores foram semelhantes para quaisquer dos métodos de avaliação testados. Estes resultados diferem em parte dos obtidos por Hengl et al. (2003), que concluíram que a amostragem estratificada foi o método mais apropriado para predizer a ocorrência de classes de solos usando mapas auxiliares e regressão logística.

Os resultados obtidos no presente trabalho demonstraram que o uso de dados independentes é o mais indicado para validar modelos preditores, tal como citado por Grinand et al. (2008) e Brus et al. (2011). Adicionalmente, ao utilizar a amostragem aleatória, a avaliação do modelo pode ser realizada com diferentes métodos de avaliação sem que se obtenha falsos valores (super ou subestimados) da acurácia. Isto pode estar associado ao fato de os valores obtidos pela validação cruzada ser calculada a partir de uma matriz de confusão baseada no uso de subconjuntos e, sempre que um subconjunto for utilizado para a avaliação, o mesmo não é utilizado para treinamento do classificador numa mesma rodada, sendo alternadamente trocados até que todos os subconjuntos sejam utilizados para avaliação. Por isso, a obtenção da acurácia pela validação cruzada resulta de uma estimativa da média das classificações e, por isso, é considerada como um indicador confiável para estimar o desempenho da predição de algoritmos supervisionados quando a amostragem for aleatória (Elkan, 2012). Adicionalmente, Steyerberg (2009) cita que a validação cruzada é baseada no método da divisão percentual e, como a subdivisão das amostras é realizada de forma aleatória, ambos os métodos de avaliação garantem uma independência dos dados permitindo a obtenção de valores da acurácia semelhantes à da concordância.

Na amostragem aleatória, os valores da acurácia geral e do índice kappa avaliados pelo método de validação aparente foram muito semelhantes aos obtidos pela avaliação com dados independentes e pela concordância. Todavia, o uso do método de validação aparente não é recomendado para avaliar a acurácia e acurácia de modelos preditores (Steyerberg, 2009) porque este método retorna valores superestimados, uma vez que todas as

comparações das predições são realizadas exatamente sobre as mesmas informações (das variáveis preditoras e dos solos) utilizadas para a geração do modelo. Desta forma, podem resultar em uma maior quantidade de acertos de classificação, porém apenas para os dados em que foram treinados (Steyerberg, 2009). Portanto, sua aplicação deve ser restrita apenas em situações em que o banco de dados seja muito pequeno a ponto de inviabilizar a partição dos dados em subconjuntos como os métodos de divisão percentual e validação cruzada (Elkan, 2012), fornecendo alguma estimativa da acurácia/acurácia da predição.

Com base nos resultados encontrados, esta pesquisa demonstrou que o esquema de amostragem totalmente aleatória resultou modelos e mapas preditores de ocorrência de solos mais acurados do que os esquemas de amostragem estratificada e proporcional à área de cada UM. Adicionalmente, os modelos gerados com dados aleatórios podem ser avaliados por quaisquer dos métodos de validação testados (validação aparente, divisão percentual, cruzada e dados independentes). Em relação às demais estratégias de amostragem, os resultados indicaram que os esquemas de amostragem influenciaram na capacidade preditiva dos modelos preditores e o método de validação usado para avaliar os modelos influencia na obtenção dos resultados da acurácia. Desta forma, a avaliação de modelos preditores deve ser realizada, sempre que possível, com reamostragem de dados independentes e que não foram usados para o treinamento dos modelos, tal como indicado por Grinand et al. (2008) e Brus et al. (2011). Em situações que o modelo preditor venha a ser gerado com dados totalmente aleatórios e que uma reamostragem para avaliação com dados independentes seja inviável, a avaliação do modelo pode ser realizada com o método de validação cruzada ou divisão percentual.

4.4 Conclusões

Os esquemas de amostragem influenciaram na quantidade de UM preditas e na acurácia dos modelos preditores, sendo que o modelo preditor de ocorrência de solos gerado com dados do esquema de amostragem aleatório simples foi o mais acurado.

A acurácia dos modelos preditores gerados com dados dos

esquemas estratificado e proporcional são superestimados quando avaliados pelos métodos de validação aparente, validação cruzada e com divisão percentual.

A avaliação de modelos preditores com dados independentes garante a obtenção de valores da acurácia semelhantes à da concordância para todos os esquemas de amostragem testados (aleatória, estratificada e proporcional).

Não houve influência do tamanho dos conjuntos de dados independentes usados na avaliação da acurácia dos modelos preditores.

5. CAPÍTULO IV - GERAÇÃO DE MAPA DE SOLOS ASSOCIANDO O CONHECIMENTO PEDOLÓGICO ÀS TÉCNICAS DO MAPEAMENTO DIGITAL DE SOLOS

5.1 Introdução

Os levantamentos de solos convencionais são a forma mais comum de caracterização e de mapeamento de solos e, em muitos casos, a única maneira pela qual a natureza altamente variável das relações solo-paisagem é catalogada (Scull et al., 2005). Nos mapeamentos convencionais, as relações solo-paisagem são estabelecidas a partir de interpretações de fotografias aéreas e de informações ambientais associadas ao conhecimento e experiência do pedólogo (Hengl et al., 2007). A representação cartográfica destas relações solo-paisagem é definida pela delimitação de unidade de mapeamento de solos (UM) (Bui, 2004). Estas UM podem ser simples, com apenas uma classe de solo, ou combinadas, com duas ou mais componentes sob a forma de associações, complexos ou grupos indiferenciados (IBGE, 2007). Semelhantemente aos mapeamentos convencionais de solos, as técnicas do mapeamento digital de solos (MDS) consistem em estabelecer as relações entre os solos e as feições da paisagem, porém, de forma mais quantitativa. Para isto, utilizam-se modelos quantitativos para inferir as variações espaciais e temporais dos solos, a partir de mapas já existentes, de observações a campo, do conhecimento pedológico e de variáveis ambientais

correlacionadas (Lagacherie & McBratney, 2007). O modelo conceitual usado no mapeamento digital de classes de solo pode ser descrito da seguinte forma (McBratney et al., 2003):

Sc = f (s, c, o, r, p, a, n)

onde Sc é a classe de solo, ***f*** é uma função empírica supervisionada ou não supervisionada, ***s*** refere-se a informações sobre o solo a partir de um mapa, banco de dados ou a partir de um conhecimento especializado, ***c*** refere-se ao clima, ***o*** refere-se a organismos, incluindo a atividade humana, ***r*** refere-se ao relevo ou topografia, ***p*** refere-se ao material de origem, **a** refere-se a idade, e ***n*** refere-se a posição espacial. Cada elemento é representado por um conjunto de uma ou mais variáveis contínuas ou categóricas, por exemplo, o ***r*** representado por elevação e declividade (Minasny & McBratney, 2007).

As técnicas preditivas totalmente diretas usadas no MDS podem não ser as mais adequadas quando usadas sem um controle da predição e de ajustes necessários a serem realizados por um pedólogo, controlando o processo de produção do mapa final (Hengl et al., 2007). Desta forma, a incorporação do conhecimento do pedólogo aos modelos preditivos, citada como conhecimento especialista ou "*expert knowledge*" por Scull et al. (2005), pode contribuir para a geração de modelos e mapas preditores de solo mais acurados. Alguns estudos (Zhu et al., 2001; Minasny e McBratney, 2007; Hengl et al., 2007) têm incorporado, como conhecimento pedológico, a distância taxonômica de solo aos modelos e, assim, demonstrado que o conhecimento e experiência do pedólogo podem ser combinados às técnicas do MDS.

A capacidade da predição de ocorrência de solos por métodos diretos, em que a partir de mapas de solos os modelos são treinados e geram diretamente um mapa digital de solos, pode ficar limitada quando as UM forem do tipo combinadas ou quanto mais complexa for a paisagem. Isto ocorre porque dentro das UM delineadas nos mapas de referência, além da complexa distribuição espacial dos solos, podem ocorrer muitas variações ambientais que prejudicam o treinamento e a capacidade preditiva dos modelos. Este tipo de procedimento de MDS baseado no treinamento de modelos a partir de um mapa de referência pode se tornar inviável na falta de mapas de solos pré-existentes. Desta forma, o uso da predição de ocorrência de classes de solos pelo uso de modelos treinados a partir de perfis de solo georreferenciados

pode ser uma alternativa para a geração de mapas digitais de classes de solos, visando mapear as áreas em que foram amostradas, bem como extrapolar a predição de ocorrência de solos para áreas ambientalmente semelhantes e que ainda não tenham sido mapeadas.

Nestes casos, ao invés de predizer a ocorrência de UM de solos a partir de mapas de solos já existentes, é realizada a predição de ocorrência de classes de solos a partir de investigações dos solos a campo ou de banco de dados já estruturados, gerando um mapa com manchas descontínuas de ocorrência de diferentes classes de solos. Sobre este mapa, o pedólogo pode exercitar e associar sua experiência e conhecimento às técnicas preditivas do MDS, para gerar um mapa de solos compostos por UM simples ou combinadas, conforme a distribuição espacial dos tipos de solos realizada pela modelagem. O objetivo deste estudo foi avaliar e comparar dois métodos de predição de ocorrência de UM (denominados de direto e indireto) para uma bacia hidrográfica do Rio Grande do Sul, utilizando técnicas de geoprocessamento e do MDS.

5.2 Material e Métodos

A área de estudo é a bacia do Rio Santo Cristo que está situada na região fisiográfica do Alto Uruguai no Estado do Rio Grande do Sul (Figura 12a). A bacia do Rio Santo Cristo apresenta uma área de 898 km^2 e abrange sete municípios da região noroeste do RS. O clima da região, segundo Köppen, é subtropical úmido (Cfa). Os solos da região são formados a partir de rochas básicas da Formação Serra Geral (Nardy et al., 2008). A geomorfologia da região se apresenta como planalto profundamente recortado pelos afluentes do Rio Uruguai, apresentando relevo suave ondulado nas proximidades das nascentes dos Rios Santo Cristo e Tuparendi e mais acidentado em direção ao Rio Uruguai, com vales profundos em V e encostas íngremes. O material de origem, o clima e a vegetação foram considerados uniformes na bacia. Devido à influência do relevo sobre formação dos solos, por condicionar os fluxos de água, os processos erosivos e a acumulação de materiais, e pela ampla disponibilidade de informações espaciais como os MDE, o relevo foi considerado como o principal fator diferenciador dos tipos de solo neste estudo.

Os atributos do terreno (elevação, declividade, direção do fluxo, acúmulo de fluxo, comprimento do fluxo, índice de umidade topográfica e curvatura) foram gerados no programa ArcGis 9.3 (ESRI, 2009) a partir do modelo digital de elevação (MDE) ASTER GDEM v.2 com tamanho de pixel de 30 m (Meyer et al., 2012). A variável distância horizontal das redes hidrográficas foi gerada a partir do arquivo vetorizado da hidrografia disponível em Hasenack & Weber (2010). A partir das correlações entre os atributos do terreno e a ocorrência dos solos foram gerados mapas digitais de solos usando dois métodos distintos. O primeiro método (denominado como método direto) consistiu na geração de um mapa preditor de ocorrência de UM, cujo modelo foi treinado com informações provenientes de um mapa convencional de solo na escala 1:50.000 (Kämpf et al., 2004). O segundo método (método indireto) consistiu em gerar um mapa de predição de ocorrência de classes de solo (mapa intermediário) a partir de informações taxonômicas de 193 perfis de solos georreferenciados, o qual foi utilizado pelo pedólogo para delinear as UM. Na Figura 13 é apresentado um fluxograma dos métodos de predição de ocorrência de solos denominados como método direto e indireto.

No método direto, o treinamento do modelo foi realizado a partir de uma amostra de 45.000 pontos distribuídos aleatoriamente na bacia. Em cada ponto foram coletados os dados das oito variáveis preditoras e das UM ocorrentes no mapa de solos convencional, na escala de 1:50.000 (Figura 12b). Os dados tabulados foram exportados para treinamento do algoritmo de árvore de classificação SimpleCart com número de elementos no nó final (M) igual a 2, no programa de mineração de dados Weka 3.6.3 (Hall et al., 2009). As regras de classificação resultantes foram implementadas em ambiente de SIG utilizando o programa ArcGis 9.3 (Esri, 2009), obtendo-se assim, o mapa de predição de ocorrência de UM gerado pelo método direto.

No método indireto, primeiramente foi gerado o mapa de predição de ocorrência de classes de solo em cada pixel. Para isto, foram utilizadas as informações taxonômicas oriundas de 193 perfis de solos georreferenciados (Kämpf et al., 2004) (Figura 12a). As classes de solos foram agrupadas ao nível de Subordem do Sistema Brasileiro de Classificação de Solos (Santos et al., 2013), por ser compatível com o tipo de levantamento de solo de reconhecimento de alta intensidade na escala de 1:50.000 (IBGE, 2007). Com

isso, totalizaram seis grupos taxonômicos, sendo 14 Cambissolos Háplicos, 13 Gleissolos Háplicos, 64 Latossolos Vermelhos, 8 Chernossolos Háplicos, 19 Neossolos Litólicos e 75 Neossolos Regolíticos. Estas observações a campo foram usadas como informação taxonômica de solos para o treinamento do modelo preditor de classes de solo.

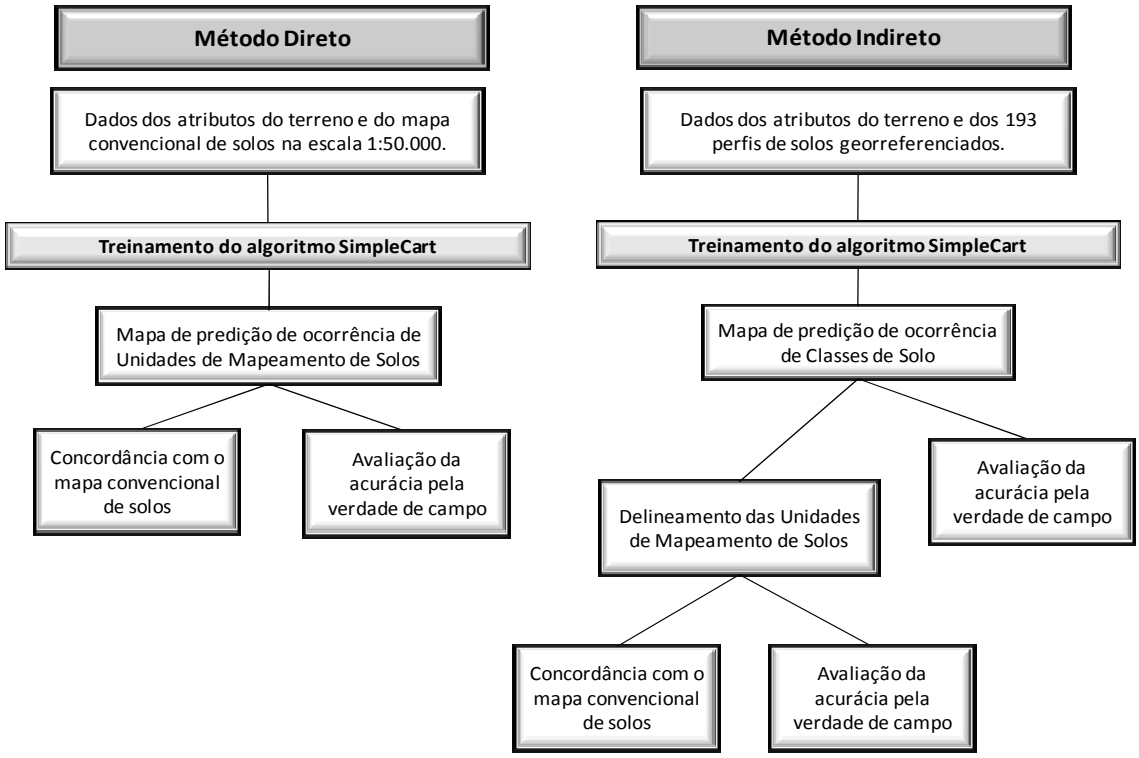

Figura 13. Fluxograma dos métodos (diretos e indiretos) usados para a predição de ocorrência de unidades de mapeamento de solos.

Nestes mesmos 193 pontos georreferenciados foram coletados em ambiente de SIG os dados dos atributos do terreno. Os dados foram tabulados e exportados para treinar o algoritmo de árvores de classificação SimpleCart com valor de M igual a 2, no programa Weka 3.6.3 (Hall et al., 2009). A partir das regras de classificação foi gerado um mapa de predição de ocorrência de classes de solo (Figura 13b), que por sua vez, foi sobreposto ao mapa com relevo sombreado derivado do MDE ASTER GDEM para o pedólogo proceder ao delineamento manual das UM. Como critério para o delineamento das UM, foram mantidas as mesmas UM e proporções das classes de solos apresentadas na legenda (Tabela 6) do mapa de solos já existente.

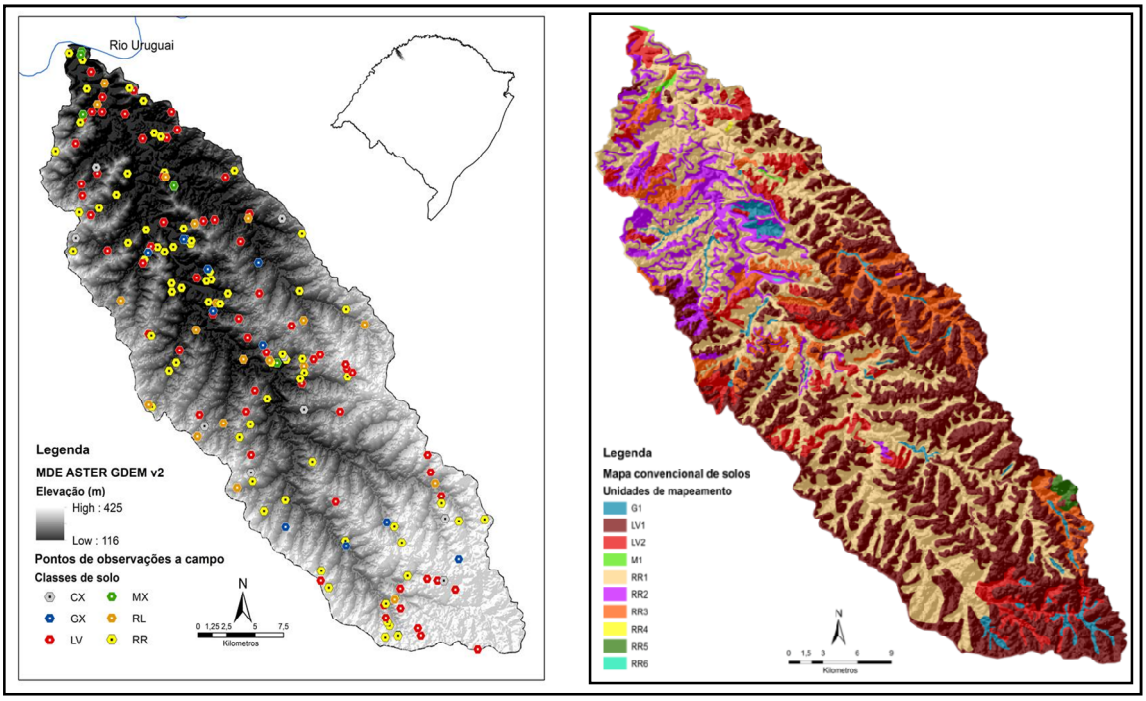

Figura 14. a) modelo digital de elevação ASTER GDEM v2 com a distribuição espacial perfis de solo, sendo CX: Cambissolo Háplico; GX: Gleissolo Háplico; LV: Latossolo Vermelho; MX: Chernossolo Háplico; RL: Neossolo Litólico; RR: Neossolo Regolítico; b) mapa convencional de solos da bacia do Rio Santo Cristo (Kämpf et al., 2004).

Tabela 6. Unidades de mapeamento de solos ocorrentes na bacia do Rio Santo Cristo (RS) (Kämpf et al., 2004).

UM*	Descrição taxonômica SBCS (Santos et al., 2013)	Proporção dos componentes (%)	Inclusões	Área (%)
G1	Gleissolo Háplico	-		2,5
LV1	Latossolo Vermelho distroférrico	-	RR, CX	38,3
LV2	Associação Latossolo Vermelho + Neossolo Regolítico	60 e 40	CX	7,9
M1	Chernossolo Háplico	-	CX	0,2
RR1	Associação Neossolo Regolítico + Cambissolo Háplico	60 e 40	RL, CX	34,8
RR2	Complexo Neossolo Regolítico + Neossolo Litólico	50 e 50	CX	8,1
RR3	Associação Neossolo Regolítico + Latossolo Vermelho	60 e 40	CX	7,9
RR4	Associação Neossolo Regolítico + Neossolo Litólico	70 e 30		0,03
RR5	Associação Neossolos Regolíticos + Cambissolos Háplicos + Latossolo Vermelho	50, 30 e 20		0,4
RR6	Associação Neossolo Regolítico + Chernossolo Háplico	60 e 40	CX	0,01

*UM: unidades de mapeamento de solos; RR: Neossolo Regolítico; CX: Cambissolo Háplico; RL: Neossolo Litólico.

A acurácia geral dos mapas convencional de solos, de predição direta de UM, de predição de classes de solo e do mapa de predição indireta de UM foi avaliada pela verdade de campo, utilizando os 193 perfis de solos georreferenciados. A concordância dos mapas preditores de UM (direto e indireto) com o mapa convencional foi avaliada por matrizes de erros (Congalton, 1991) obtendo-se a concordância geral, a acurácia do mapeador, os erros de omissão e o índice Kappa (Cohen, 1960).

5.3 Resultados e Discussão

A avaliação pela verdade de campo do mapa convencional de solos resultou em uma acurácia geral de 79,8 % (Tabela 7). As UM que ocorrem em áreas mais extensas LV1, LV2, RR1 e RR3 apresentaram as maiores porcentagens de acertos com valores iguais ou maiores que 78,8 %, sendo as UM LV1 e LV2 as mais acuradas. Estas duas UM representam, principalmente, a distribuição dos Latossolos Vermelhos que ocorrem em extensas e contínuas áreas da bacia, cujas características do relevo se apresentam com menores variações do que nas áreas de ocorrência das UM RR1, RR2 e RR3 que representam como principais componentes os solos menos desenvolvidos da bacia como os Cambissolos, Neossolos Regolíticos e Neossolos Litólicos. Nas demais UM avaliadas pela verdade de campo (RR2, M1 e G1), foram encontradas menores porcentagens de acertos, porém, com valor igual ou maior que 50,0 %.

O treinamento do algoritmo SimpleCart usando as informações do mapa convencional de solos e dos atributos do terreno resultou uma árvore de classificação com tamanho de 145 e número de folhas igual a 73. Nesta árvore, os atributos do terreno que melhor explicaram a ocorrência de UM foram a elevação, a declividade, o comprimento de fluxo e a distância horizontal das redes hidrográficas. A árvore de classificação gerada a partir das informações taxonômicas dos 193 perfis de solos e dos atributos do terreno apresentou um menor tamanho, com 99 subdivisões e o número de folhas foi igual a 50. Os atributos do terreno que melhor explicaram a ocorrência das classes de solos foram a distância horizontal das redes hidrográficas, a elevação, a declividade e o comprimento do fluxo.

Tabela 7. Resultados da avaliação da acurácia pela verdade de campo no mapa de solos convencional, no mapa de predição de classes de solo e nos mapas de predição direta e indireta de unidades de mapeamento.

UM	Mapa convencional	Mapa preditor direto	Mapa preditor indireto	CS	Mapa preditor de CS
	---------- acertos % ----------				--acertos %--
G1	50,0	-	71,4	CX	15,6
LV1	86,8	92,9	89,5	GX	54,5
LV2	84,0	66,7	92,3	LV	60,5
M1	66,7	-	80,0	MX	80,0
RR1	78,8	62,1	78,4	RL	57,1
RR2	66,7	100,0	50,0	RR	66,1
RR3	83,3	-	87,0	-	-
RR4	-	-	-	-	-
RR5	-	-	-	-	-
RR6	-	-	-	-	-
AG (%)	**79,8**	**73,6**	**78,8**	**-**	**54,9**

*UM: unidades de mapeamento de solos; CS: classes de solo; AG: acurácia geral; CX: Cambissolo Háplico; GX: Gleissolo Háplico; LV: Latossolo Vermelho; MX: Chernossolo Háplico; RL: Neossolo Litólico; RR: Neossolo Regolítico.

O mapa de predição direta de UM (Figura 13a) estimou seis UM (G1, LV1, LV2, RR1, RR2 e RR3) com uma acurácia geral de 73,6 % avaliada pela verdade de campo (Tabela 7). A UM mais extensa (LV1) foi predita em áreas com relevo plano a suave ondulado apresentando a segunda maior acurácia com 92,9 % de acertos. Nas áreas com relevo ondulado a forte ondulado foram preditas as UM RR1, RR2 e RR3. A predição da segunda UM mais extensa (RR1) foi generalizada por diversas regiões da bacia, ocupando desde áreas de encosta até o fundo dos vales e apresentou uma acurácia de 62,1 %.

A concordância do mapa de predição direto de UM com o mapa convencional de solos foi de 63,0 % e um índice Kappa de 0,43 (Tabela 8). Estes resultados são similares aos obtidos por Giasson et al. (2011), que encontraram valores de concordância geral entre 52,1 % e 71,7 % e índice Kappa de 0,33 e 0,57. No mapa de predição direta, as UM LV1 e RR1 permaneceram como as UM mais extensas ocupando 43,9 % e de 39,1 % do total da bacia e foram preditas de forma semelhante ao mapa convencional resultando nos maiores valores de acurácia do mapeador. E em relação às áreas ocupadas por estas UM no mapa convencional, 80,9 % foram igualmente classificados no mapa preditor como UM LV1 e 69,3 % foram corretamente classificados como RR1. Os erros de classificação ocorreram em geral, pela mútua confusão da predição das UM LV1 e UM RR1, e isto pode estar

relacionado ao fato das áreas de ocorrência destas UM no mapa convencional serem adjacentes e alternadas em praticamente toda a bacia.

Tabela 8. Matriz de erros comparando o mapa convencional de solos e o mapa de predição de ocorrência de UM gerado pelo método direto.

		Mapa de referência										
	UM*	UM										Total (%)
		G1	LV1	LV2	M1	RR1	RR2	RR3	RR4	RR5	RR6	
Mapa produzido (direto)	G1	0,4	0,1	0,2	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,7
	LV1	0,5	31,0	2,7	0,0	6,4	0,7	2,3	0,0	0,3	0,0	43,9
	LV2	0,3	0,5	1,6	0,0	0,7	0,2	0,1	0,0	0,0	0,0	3,4
	M1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	RR1	1,0	5,4	2,7	0,2	24,1	3,3	2,3	0,0	0,1	0,0	39,1
	RR2	0,1	0,6	0,5	0,0	2,4	3,7	0,7	0,0	0,0	0,0	8,0
	RR3	0,1	0,7	0,2	0,0	1,2	0,2	2,5	0,0	0,0	0,0	4,9
	RR4	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	RR5	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	RR6	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	Total	2,4	38,3	7,9	0,2	34,8	8,1	7,9	0,0	0,4	0,0	100,0
	AM (%)	17,7	80,9	20,3	0,0	69,3	45,7	31,6	0,0	0,0	0,0	
	AG (%)	63,0										
	Kappa	0,46										

*UM = unidades de mapeamento de solos; AG = acurácia geral; AM = acurácia do mapeador.

Para a obtenção do mapa contendo as UM delineadas pelo pedólogo, foi gerado primeiramente, como produto intermediário, um mapa de predição de ocorrência de classes de solo (Figura 13b). A predição das seis classes de solo encontradas nas observações a campo resultou uma distribuição espacial destes solos com uma acurácia avaliada pela verdade de campo de 54,9 % (Tabela 7). Com exceção dos Cambissolos, a porcentagem de acertos na predição dos tipos de solo variou de 54,5 % a 80 %, valores estes, são semelhantes aos encontrados por Bui& Moran (2003) e Minasny & McBratney (2007) cujo percentual de predição correta de cada classe de solo variou entre 50,0 a 80 %. Em estudo de predição de classes de solo por árvore de classificação e um algoritmo genérico, Nelson & Odeh (2009) utilizando informações taxonômicas de 3.875 perfis de solo encontraram valores de índice Kappa entre 0,07 a 0,37 e acurácia com 10 % a 53 % de acertos de classificação. Adhikari et al. (2014) ao utilizarem 1.171 perfis de solo como informação pedológica de referência, a predição de tipos de solos por árvore de decisão resultou valores de acurácia de 51,0 % a 60,0 % na validação do modelo. Ao validar mapas de predição de classes de solo utilizando pontos a

campo, Hengl et al. (2007) encontraram valor máximo de classificações corretas de 36,7 % e Roecker et al. (2010) encontraram uma acurácia geral de 49,0 % ao predizer a ocorrência de subgrupos de solos.

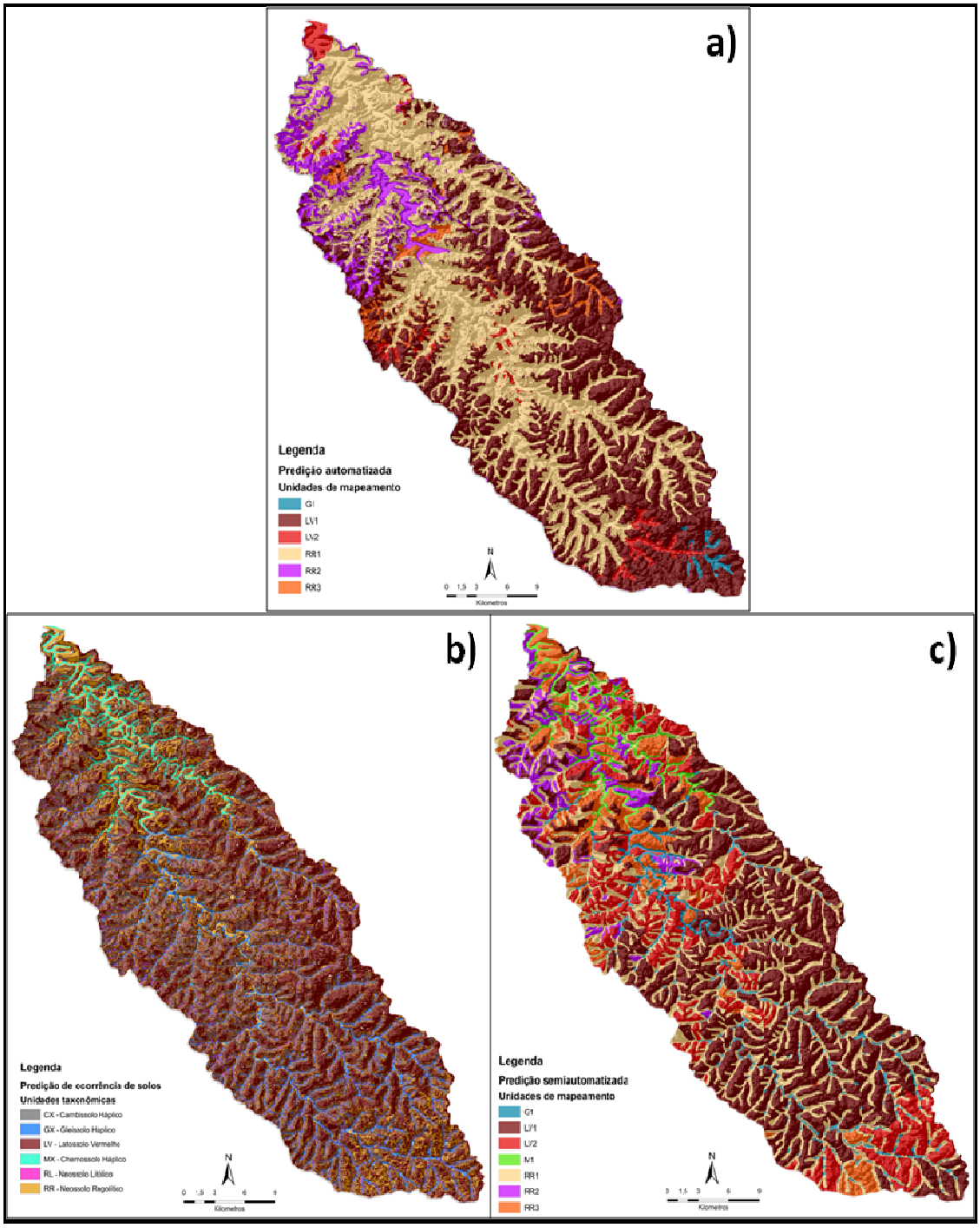

Figura 15. a) mapa de unidades de mapeamento de solos gerado pelo método direto; b) mapa intermediário de ocorrência de classes de solos; c) mapa de predição de ocorrência de unidades de mapeamento de solos delineadas a partir do mapa de predição de classes de solo (método indireto).

Neste mapa de classes de solos, a distribuição espacial dos Latossolos Vermelhos foi predita para as áreas com altitudes superiores a 239 m, ocupando extensas áreas de relevo suave-ondulado a ondulado. A classe de solo Neossolo Regolítico foi predita em áreas do sudeste da bacia com relevo ondulado e ocupando, preferencialmente, os terços inferiores de encosta com declividade maior que 20 % nas demais regiões. A predição da classe de solo Neossolo Litólico na porção mais a noroeste da bacia ocorreu nas áreas mais altas (acima de 170 m) com relevo fortemente ondulado a montanhoso e nas demais regiões foi predita desde as áreas mais altas, passando pelas posições de terço médio das encostas, até o fundo dos vales encaixados em forma de V. A maior parte da predição da classe de solo Cambissolo foi distribuída nas áreas com relevo mais acidentado da bacia, geralmente acompanhando a ocorrência das classes de solo Neossolo Regolítico e Neossolo Litólico, formando um mosaico com variadas proporções destes três tipos de solo, principalmente, nas áreas de encostas mais declivosas. Nas áreas de fundo de vales planos da bacia foram preditas a ocorrência de Gleissolo Háplico nas porções central e sudeste da bacia e os Chernossolos Háplicos foram preditos na região noroeste da bacia.

Estes mapas de predição de ocorrência de classes de solo (Figura 13b) mostram a variação de ocorrência dos tipos de solos para cada pixel de 30 x 30 m, porém a distribuição espacial dos solos é cartograficamente inadequada e pouco legível, o que pode dificultar a interpretação da ocorrência dos solos de uma dada região devido a grande variabilidade espacial (Roecker et al., 2010). Assim, este mapa de distribuição de classes de solos foi redesenhado através do delineamento manual de UM de solos, quando necessário criando UM combinadas como associações de solos, devido à impossibilidade de gerar UM simples para todas as classes de solos.

No mapa gerado pelo delineamento manual de UM sobre o mapa de classes de solos preditas foram identificadas e delineadas 7 UM (Figura 13c). A UM LV1 foi delimitada, principalmente, em cotas altimétricas acima de 240 m e em áreas com relevo plano a ondulado, ocupando extensas áreas desde o sudeste da bacia até áreas próximas da porção central. Os maiores delineamentos da UM LV2 foram realizados a sudeste e na porção central e, em menor proporção, em áreas situadas a noroeste da bacia. Estas UM (LV1 e

LV2) foram delineadas em sua maioria, nas áreas da bacia com relevo mais suave o que favorecem os processos de percolação de água no perfil e a formação de solos mais profundos, tal como os Latossolos Vermelho presentes em ambas as legendas. As UM RR1, RR2 e RR3 foram delineadas nas áreas mais íngremes da bacia e, geralmente, ocupando os terços médios e inferiores de encostas até o fundo de vales de menor altitude, e com menor expressividade nas áreas de topo dos morros. As UM simples (G1 e M1) foram delineadas nos fundos dos vales, onde foi predita a ocorrência de Gleissolos Háplicos e Chernossolos Háplicos.

As avaliações da acurácia pela verdade de campo no mapa de predição indireta de UM resultaram uma acurácia geral de 78,8 % (Tabela 7). Este valor foi muito semelhante ao encontrado no mapa de solos convencional (79,8 %) e levemente superior ao mapa de predição direta de UM (73,6 %). Os resultados da verdade de campo demonstraram que o mapa de predição indireta de UM apresentou porcentagens de acertos iguais ou maiores a 50% para todas as UM e possibilitou ao pedólogo delimitar as UM G1, LV1, LV2, M1, RR1 e a RR3 com acurácia igual ou maior do que no mapa de solos convencional. Portanto, além das UM mais representativas da bacia do Rio Santo Cristo, que também foram preditas pelo método direto, este método possibilitou delinear duas das UM menos extensas da bacia, as UM simples G1 e M1, com respectivas acurácias de 71,4 % e 80,0 %. Isto pode ser considerado vantajoso, pois a reprodutibilidade de UM menos extensas pelos modelos preditores diretos é uma dificuldade comumente encontrada e relatada em diversos trabalhos de MDS (Giasson et al., 2011; ten Caten et al., 2011).

A comparação do mapa de predição indireta de UM com o mapa convencional apresentou uma concordância geral de 42,0 % (Tabela 9). Os maiores valores de acurácia do mapeador, foram encontrados na UM LV1 com 64,4 %, seguido da UM M1 com 41,1 % e da UM RR1 com 35,5 % das classificações desta UM foram concordantes com o mapa original de solos. Tanto no mapa de referência como no mapa com as UM delineadas a partir da predição de classe de solo, as UM LV1 e RR1 foram as mais extensas, embora apresentem diferentes percentuais de área ocupada em cada mapa. Sendo que a UM LV1 foi delineada de forma concordante em 64,4 % com o mapa convencional de solos, cujos erros de omissão desta UM são provenientes,

principalmente, do delineamento das UM LV2 (18,3 %) e da RR1 (11,5 %) em áreas mapeadas originalmente como UM LV1. A UM RR1 apresentou uma concordância de 35,5 % com o mapa de solos convencional, sendo que na área ocupada pela RR1 no mapa convencional, 26,7 % foram delineadas como LV1, 10,9 % como LV2 e 10,3 % como RR3. Assim como no mapa gerado pela predição direta de UM, estas confusões de classificação entre as UM mais extensas da bacia, a LV1 e a RR1, estão relacionadas ao fato destas UM serem adjacentes em praticamente todas as regiões da bacia e também devido aos atributos do terreno avaliados se apresentaram de forma muito semelhantes nas áreas próximas dos limites destas duas UM. Assim, ao utilizar somente as informações das variáveis do terreno, a distinção de ocorrência destas UM foi confundida em relação ao mapa de solos convencional. Contudo, as confusões de classificações de algumas UM podem ser consideradas menos importantes, por representarem cartograficamente características similares da paisagem, bem como, a distribuição espacial das mesmas classe de solo, tal como ocorre na UM RR1, em que as classes de solo Neossolo Regolítico e Cambissolo Háplico são também inclusões na UM LV1.

Tabela 9. Matriz de erros comparando o mapa de predição de unidades de mapeamento de solos gerado pelo método indireto e o mapa convencional de solos.

		Mapa de referência										
	UM*	**UM**										**Total (%)**
		G	**LV1**	**LV2**	**M1**	**RR1**	**RR2**	**RR3**	**RR4**	**RR5**	**RR6**	
Mapa produzido (indireto)	**G1**	0,5	0,3	0,3	0,0	2,0	0,3	0,5	0,0	0,0	0,0	3,9
	LV1	0,3	24,7	1,6	0,0	9,3	1,5	2,4	0,0	0,3	0,0	40,1
	LV2	0,2	7,0	2,3	0,0	3,8	0,8	1,0	0,0	0,0	0,0	15,1
	M1	0,1	0,0	0,2	0,1	2,0	0,8	0,0	0,0	0,0	0,0	3,2
	RR1	0,9	4,4	1,9	0,0	12,4	2,0	2,8	0,0	0,1	0,0	24,3
	RR2	0,2	0,2	0,6	0,0	1,7	1,2	0,6	0,0	0,0	0,0	4,6
	RR3	0,2	1,8	0,9	0,1	3,6	1,5	0,6	0,0	0,0	0,0	8,7
	RR4	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	RR5	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	RR6	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	Total	2,4	38,3	7,9	0,2	34,8	8,1	7,9	0,0	0,4	0,0	100,0
	AM (%)	21,2	64,4	29,4	41,1	35,5	14,4	7,1	0,0	0,0	0,0	
	AG (%)	42,0										
	Kappa	0,2										

*UM = unidades de mapeamento de solos; AG = acurácia geral; AM = acurácia do mapeador.

Com base nos resultados encontrados, podemos concluir que o

mapa de predição direta de UM apresentou uma concordância maior com mapa convencional de solos (63,0 %) do que o mapa de predição indireta de UM (42,0 %). Porém, a avaliação da acurácia pela verdade de campo indicou que a integração do conhecimento do pedólogo na geração do mapa de solos resultou um mapa de predição indireta de UM mais acurado (acurácia de 78,9 %) do que o mapa gerado pelo método direto (acurácia de 73,0% e quase tão acurado quanto o mapa de solos convencional (acurácia de 80,0%). Dentre os poucos estudos comparando a acurácia de mapas gerados pelas técnicas do MDS e convencionais, Skidmore et al. (1996) encontraram acurácia de 74 % no mapa convencional e de 70,0 % no mapa gerado com classificador supervisionado Bayesiano. Zhu et al. (2001) relatam resultados concordantes com os encontrados neste estudo, em que os mapas de solos produzidos pelo levantamento convencional do solo foram menos acurados (61,0 % e 67,0 % de acerto) do que os mapas de solos produzidos pelo modelo de inferência de relações solo-paisagem supervisionado pelo pedólogo (81,0 % e 84,0 % de acerto). Este estudo demonstrou que, com o acréscimo do conhecimento do pedólogo às técnicas quantitativas de predição de classe de solo, foi possível delinear as UM de forma tão acurada quanto no mapa convencional de solos, com vantagem em relação ao método de predição direta de UM por possibilitar a geração de mapa digital de solos sem a necessidade de mapas pedológicos para o treinamento de modelos preditores.

5.4 Conclusões

O mapa de predição de ocorrência de unidades de mapeamento de solo gerado pelo método direto apresentou maior concordância com o mapa convencional de solos do que o mapa predito pelo método indireto.

O mapa de unidades de mapeamento de solo gerado com o delineamento de UM pelo pedólogo apresentou a maior acurácia avaliada pela verdade de campo.

A associação do conhecimento do pedólogo à predição de classes de solo pelas técnicas do mapeamento digital de solos é um método especialmente útil na falta de mapas pedológicos de referência para o treinamento dos modelos preditores.

6. CONCLUSÕES GERAIS

A presente tese avaliou e comparou o uso de diferentes métodos e materiais para o desenvolvimento e aplicação das técnicas do mapeamento digital de solos (MDS), considerando mais especificamente a predição de ocorrência de classes de solo. Nesta pesquisa foram apresentados e discutidos quatro estudos, sendo uma Revisão Bibliográfica sobre as árvores de decisão e seu uso no MDS, e três estudos de predição de ocorrência de classes de solos para duas regiões do estado do Rio Grande do Sul.

A Revisão Bibliográfica possibilitou compreender como as árvores de decisão particionam recursivamente o conjunto de dados, em que um problema mais complexo é decomposto em subproblemas mais simples, até que os subconjuntos apresentem uma única classe ou valor predito. Desta forma, os modelos de árvores de decisão se apresentam em uma estrutura hierárquica que se desenvolve da raiz para as folhas. As ávores de decisão têm sido consideradas eficientes para predizer a ocorrência de solos, pois são considerados algoritmos robustos, lidam com variáveis tanto categóricas como numéricas e relacionam a ocorrência dos solos com as características da paisagem de forma análoga às correlações mentais realizadas pelo pedólogo (aprendizado humano indutivo). Além de as pesquisas que reportem o uso de árvores de decisão no MDS serem recentes no Brasil e que poucos são os profissionais que detêm conhecimento para sua utilização, mais estudos com esta técnica serão úteis para diminuir possíveis limitações preditivas, tais como o emprego de diferentes estratégias para geração de árvores de menor

tamanho e com maior acurácia, bem como a integração do conhecimento do pedólogo para refinar e ajustar os dados de solos e de variáveis ambientais a serem usados para treinamento dos modelos.

No estudo realizado para o município de Dois Irmãos (Capitulo II) foi avaliado e comparado o uso de diferentes fontes e resoluções espaciais de MDE para derivar atributos do terreno correlacionados com a ocorrência dos solos. Foram utilizados os MDE oriundos do ASTER GDEM v2 (30 m), do TOPODATA (30 m), do SRTM v4.1 (90 m), do Brasil em Relevo (90 m) e os MDE gerados a partir das curvas de nível nas resoluções de 30 m (CN30) e 90 m (CN90). Os resultados deste estudo demonstraram que, com o uso de árvores de decisão, os atributos do terreno que melhor explicaram a ocorrência dos solos foram a elevação, a declividade, o comprimento de fluxo e a orientação das vertentes. Adicionalmente, o tipo e a resolução espacial dos MDE influenciaram na acurácia dos modelos preditores de ocorrência de solos. Por se tratar de uma área com predomínio de relevo plano a suave ondulado, os MDE com resolução espacial de 90 m (SRTM v4.1 e CN90) apresentaram menores oscilações nos valores dos atributos do terreno dentro de cada UM, o que possibilitou gerar modelos preditores de ocorrência de solos mais acurados e com maior número de UM de solo preditas. Desta forma, na predição de ocorrência de classes de solos para áreas que apresentam relevo mais suave usando informações dos solos extraídas de mapas pedológicos na escala de 1:20.000, os MDE com resolução espacial de 90 m demonstraram ser os mais indicados para derivar os atributos do terreno a serem usados como variáveis preditoras de ocorrência de solos. Todavia, mais estudos a respeito da influência dos diferentes tipos e resoluções espaciais dos MDE sobre a predição de ocorrência de classes de solo se tornam necessários, principalmente, se forem realizadas em áreas com predomínio de fases de relevo mais complexas ou que estejam cobertas por mapas de solos mais detalhados, para que sejam definidos os tipos dos MDE a serem usados no MDS conforme a complexidade das características da paisagem e a distribuição espacial dos solos.

No MDS, as classes e UM de solos pouco extensas e pouco representativas não são preditas quando amostradas de forma totalmente aleatória. No Capítulo III foram avaliados e comparados três esquemas de

amostragem usados para coletar dados usados no treinamento de modelos preditores de ocorrência de solos para a microbacia do Rio Santo Cristo. Foram testados o esquema de amostragem aleatório simples, aleatório proporcional à área ocupada por cada UM e estratificado pelo número de UM. Adicionalmente, foram testados diferentes métodos de avaliação da acurácia dos modelos preditores. O modelo preditor e o mapa digital de ocorrência de solos gerados com os dados da amostragem aleatória, embora não tenham predito todas as UM como na amostragem estratificada, foram os mais acurados por distribuir os solos na paisagem de forma mais semelhante ao mapa convencional de solos. Quando os dados para treinamento são amostrados de forma totalmente aleatória, a avaliação da acurácia dos modelos preditores de ocorrência de solos, gerados por árvore de decisão, pode ser realizada por quaisquer métodos testados (validação aparente, validação cruzada, validação com divisão percentual ou com dados independentes). As estimativas da acurácia dos modelos preditores gerados com dados provenientes da amostragem estratificada ou proporcional foram superestimadas quando avaliados pelos métodos de validação aparente, validação cruzada e com divisão percentual.

Desta forma, este estudo demonstrou que nos modelos preditores de ocorrência de solos construídos por árvore de decisão a partir de informações extraídas de um mapa convencional de solos e de mapas auxiliares, o esquema de amostragem aleatório mostrou-se o mais vantajoso para o MDS por gerar os modelos preditores e mapas de ocorrência de solos mais acurados e sem que a acurácia dos modelos fosse superestimada. Ademais, a avaliação da acurácia de modelos preditores gerados com dados oriundos das amostragens estratificada e proporcional deve ser realizada, exclusivamente, com dados independentes e distintos daqueles usados para treinamento dos modelos. Assim, estudos de predição de ocorrência de classes de solos que utilizem informações dos solos extraídas de mapas pedológicos já existentes para o treinamento dos modelos podem adotar a amostragem aleatória simples para garantir a geração de modelos preditores e mapas digitais de classes de solos mais acurados. Todavia, no Brasil ainda são recentes os trabalhos com MDS e, por isso, outros esquemas de amostragem baseados na distribuição espacial dos solos ou nas formas do relevo podem contribuir ainda mais para a seleção e definição da amostragem a ser usada

para estudos de predição de ocorrência de solos brasileiros.

No Capítulo IV foi proposto, avaliado e discutido um método indireto para a geração de mapas digitais de ocorrência de UM de solos. Devido à escassez de mapas pedológicos, a predição de ocorrência de solos pelas técnicas do MDS, baseada no treinamento de modelos a partir de um mapa de referência (método direto), pode se tornar inviável na falta de mapas de solos pré-existentes. Assim, o método proposto neste estudo (indireto) consistiu em delinear as UM de solos usando um mapa digital de ocorrência de tipos de solos que foi gerado através do treinamento de árvore de decisão, com dados de atributos do terreno e de informações taxonômicas de perfis de solos georreferenciados. Os resultados deste estudo demonstraram que o mapa de predição de ocorrência de UM de solo gerado pelo método direto foi mais concordante com o mapa convencional de solos do que o mapa digital de UM de solo, porém o delineamento manual das UM (método indireto) possibilitou gerar um mapa com a maior acurácia avaliada pela verdade de campo. Desta forma, a associação do conhecimento do pedólogo à predição de classes de solo pelas técnicas do MDS demonstrou ser um método especialmente útil para elaborar mapas digitais de solos na falta de mapas pedológicos de referência a serem usados para o treinamento dos modelos preditores.

Neste estudo, os tipos de solos encontrados em 193 perfis de solos georreferenciados foram classificados ao nível de Subordem do Sistema Brasileiro de Classificação de Solos. Isto resultou em uma densidade de 0,0021 observação por hectare, valor este, menor que o recomendado para o tipo de levantamento semidetalhado na escala de 1:50.000 que é, na média, de 0,02 a 0,20 observação por hectare. Assim, além do método indireto possibilitar a geração de mapas de UM de solos associando o conhecimento do pedólogo às técnicas preditivas do MDS, uma menor quantidade de observações a campo foi necessária para inferir a distribuição dos solos na paisagem, de forma a resultar um mapa de UM tão acurado quanto ao mapa convencional e mais acurado que o mapa gerado pelo método direto. O método proposto neste estudo pode vir a impulsionar a aplicação das técnicas do MDS na execução de novos projetos de levantamentos, bem como para aumentar o detalhamento de mapas convencionais de solos utilizando dados já previamente levantados. Todavia, mais estudos ainda se tornam necessários para definição dos

métodos visando a obtenção de mapas digitais de solos tão acurados quanto os obtidos neste estudo. Para isso, a utilização de outras variáveis ambientais como a geologia, a vegetação, a diferenciação das formas do relevo, a classificação dos solos em níveis hierárquicos inferiores (grande grupo e sub grupo), dados de sensores remotos e proximais, entre outros poderiam vir, eventualmente, gerar mapas digitais de solos mais acurados.

Para realizar a predição de classes ou de UM de solos por correlação espacial com fatores de formação dos solos, a utilização de árvores de classificação tem sido comumente encontrada na literatura científica. No atual estudo foi utilizado o algoritmo SimpleCart no programa de mineração de dados Weka, porém, outros algoritmos de aprendizagem supervisionada e processados em outros programas estatísticos, como o R, poderiam resultar modelos preditores e mapas digitais mais acurados. Atualmente, a utilização do programa estatístico R vem recebendo uma maior atenção para o MDS, que através do consórcio *GlobalSoilMap,* vêm sendo desenvolvidos novos algoritmos e funções para a predição de classes e de propriedades do solo. Além disso, o pacote estatístico R permite a visualização espacial da predição de solos no próprio programa, bem como realizar a comunicação dos dados entre o programa estatístico e o programa de SIG de código aberto de forma mais rápida e eficiente para a geração dos mapas digitais de solos. No Brasil, os estudos com MDS têm sido realizados, em sua maioria, com treinamento de algoritmos no programa de mineração Weka e implementação dos modelos no programa ArcGis para a geração dos mapas digitais. Assim, estudos futuros com MDS podem se beneficiar do uso dos pacotes estatísticos e algoritmos de aprendizagem em máquina desenvolvidos no R e direcionados para o MDS

Mais recentemente, diversos pesquisadores vêm desenvolvendo métodos e algoritmos para gerar, e também extrapolar para áreas na mapeadas, mapas pedológicos mais detalhados do que os já existentes pelo uso da desagregação de polígonos. Como os mapas convencionais de solos não representam a distribuição individual das classes de solo, mas representam a ocorrência dos solos delimitados por polígonos, estes mapas apresentam algumas limitações, como por exemplo, a inviabilidade de identificar a ocorrência de uma determinada classe de solo dentro de uma UM. Assim, a desagregação de polígonos de solos poderia resultar na geração de

mapas pedológicos mais detalhados ao separar as classes de solo constituintes das UM de solos e mapeá-las individualmente, obtendo-se manchas de solos mais homogêneas.

Os resultados das pesquisas com MDS poderão ser pouco úteis, se não forem realmente aplicadas a novos projetos de levantamento de solos ou utilizadas para aumentar o detalhamento dos mapas já existentes. Neste sentido, há ainda uma carência de estudos voltados para definir a viabilidade de uso do MDS avaliando e comparando o tempo e os custos necessários para realizar um mapeamento de solos com uso das técnicas preditivas em relação ao uso dos métodos convencionais de mapeamento de solos. Por isso, o desenvolvimento de pesquisas que avaliem e comparem as vantagens e desvantagens de ambos os métodos poderiam permitir não só a acurácia e erros dos mapeamentos, mas também poderia se obter e comparar o tempo despendido e os custos envolvidos para cada método empregado, a fim de demonstrar a viabilidade econômica, ou não, do uso das técnicas do MDS.

7. REFERÊNCIAS BIBLIOGRÁFICAS

ADHIKARI, K. et al. Constructing a soil class map of Denmark based on the FAO legend using digital techniques. **Geoderma**, Amsterdam, v. 214-215, p.101-113, 2014.

BASGALUPP, M.P. **LEGAL-Tree:** um algoritmo genético multi-objetivo lexicográfico para indução de árvores de decisão. 2010. 116 f. Tese (Doutorado) – Pós-Graduação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2010.

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

BOU KHEIR, R. et al. Predictive mapping of soil organic carbon in wet cultivated lands using classification-tree based models: the case study of Denmark. **Journal of Environmental Management**, London, v.91, n.5, p.1150–1160, 2010.

BREIMAN, L. et al. **Classification and regression trees**. Califórnia: Wadsworth & Brooks/Cole Advanced Books & Software, 1984. 368p.

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

BRUNGARD, C.W.; BOETTINGER, J.L. Application of conditioned Latin hypercube sampling in arid Rangelands in Utah, USA. In: BOETTINGER, J.L. et al. (Ed.).**Digital soil mapping:** bridging research, environmental application, and operation. Springer, Dordrecht, 2010. p. 67-75.

BRUS, D.J., de GRUIJTER, J.J. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). **Geoderma**, Amsterdam, v.80, p.1–59, 1997.

BRUS, D.J.;KEMPEN, B.; HEUVELINK, G.B.M. Sampling for validation of digital

soil maps. **European Journal of Soil Science**, Oxford, v.62, p.394–407, 2011.

BUI, E. N. Soil survey as a knowledge system. **Geoderma**, Amsterdam, v.120, p.17–26, 2004.

BUI, E.N.; MORAN, C.J. A strategy to fill gaps in soil survey over large spatial extents: an example from the Murray–Darling basin of Australia. **Geoderma**, Amsterdam, v.111, n.1–2, p.21–44, 2003.

CAVAZZI, S. et al. Are fine resolution digital elevation models always the best choice in digital soil mapping? **Geoderma**, Amsterdam, v.195-196, p.111-121, 2013.

CHAGAS, C.S. et al. Avaliação de modelos digitais de elevação para aplicação em um mapeamento digital de solos. **Revista Brasileira de Engenharia Agrícola e Ambiental**, Campina Grande, v.14, p.218-226, 2010.

CHAPELLE, O.; SCHÖLKOPF, B.; ZIEN, A. **Semi-Supervised Learning**. Cambridge, Massachusetts, EUA:MIT Press, 2006. 43p.

CHATFIELD, C. Model uncertainty, data mining, and statistical inference (with discussion).**Journal of the Royal Statistical Society**, London, v.158, p.419-466, 1995.

COHEN, J. A coefficient of agreement for nominal scales. **Journal of Educational Measurement**, Washington, v.20,p.37-46, 1960.

CONGALTON, R.G. A review of assessing the accuracy of classification of remotely sensed data. **Remote Sensing Environmental**, New York, v.37,p.35-46, 1991.

CRIVELENTI, R. C. et al.Mineração de dados para inferência de relações solo-paisagem em mapeamentos digitais de solo. **Pesquisa Agropecuária Brasileira**, Brasília, v.44, n.12, p.1707-1715, 2009.

DU, W.; ZHAN, Z. Building Decision Tree Classifier on Private Data. In:CLIFTON, C.; ESTIVILL-CASTRO, V. (Ed.). **IEEE ICDM Workshop on Privacy, Security and Data Mining** (PSDM 2002), Maebashi, Japão, p.1-8, 2002.

ELKAN, C. **Evaluating classifiers**. San Diego, California: Department of computer Science and engineering, University of California, 2012. 25p.

ESRI. Environmental Systems Research Institute, Inc. (ESRI). **ArcGIS, Professional GIS for the desktop**, versão 9.3.1 CA. 2009.

FLORINSKY, I. V. Digital Terrain Analysis in Soil Science and Geology. Amsterdam: **Elsevier/Academic Press**, 2012. 379p.

FRIEDMAN, J.H. Greedy function approximation: a gradient boosting machine. **Annalsof Statistics,** Philadelphia, EUA, v.29, n.5, 1189–1232, 2001.

GEOBANK – CPRM. Mapa geológico do estado do Rio Grande do Sul. **Serviço**

Geológico do Brasil, 2014. Disponível em: <http://geobank.sa.cprm.gov.br/pls/publico/. Acesso em: 12 abr. 2014.

GIASSON, E. et al. Digital soil mapping using multiple logistic regression on terrain parameters in southern Brazil. **Scientia Agricola**, Piracicaba, v.63, p.262-268, 2006.

GIASSON, E.; HARTEMINK, A.E.; TORNQUIST, C.G.; TESKE, R.; BAGATINI, T. Avaliação de cinco algoritmos de árvores de decisão e três tipos de modelos digitais de elevação para mapeamento digital de solos a nível semidetalhado na Bacia do Lageado Grande, RS, Brasil. **Ciência Rural**, Santa Maria v.43,n.11,p.1967-1973, 2013.

GIASSON, E. et al.Decision trees for digital soil mapping on subtropical basaltic steeplands. **Scientia Agrícola**, Piracicaba, v.68, n.2, p.167-174, 2011.

GRINAND, C. et al. Extrapolating regional soil landscapes from an existing soil map: sampling intensity, validation procedures, and integration of spatial context. **Geoderma**, Amsterdam, v.143, p.180–190, 2008.

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

GUTH, P.L. **Geomorphometric comparison of ASTER GDEM and SRTM**. A special joint symposium of ISPRS Technical Commission IV & AutoCarto in conjunction with ASPRS/CaGIS, 2010.

HALL, M. et al. The WEKA data mining software: an update**. SIGKDD Explorations,** [S.l.], v.11, n.1, p.10-18, 2009.

HARTEMINK, A. E.et al. GlobalSoilMap.net—a new digital soil map of the world. In: BOETTINGER, J. L. et al. (Ed.).**Digital soil mapping:** bridging research, environmental application, and operation. Dordrecht: Springer Science, 2010.p. 423–428

HASENACK, H.; WEBER, E. **Base cartográfica vetorial contínua do Rio Grande do Sul - escala 1:50.000**. Porto Alegre: UFRGS- IB- Centro de Ecologia, 2010. (Série Geoprocessamento, 3) DVD-ROM

HENDERSON, B.L. et al. Australia-wide predictions of soil properties using decision trees. **Geoderma**, Amsterdam, v.124,p.383–398, 2005.

HENGL, T.; ROSSITER, D. G.; STEIN, A. Soil sampling strategies for spatial prediction by correlation with auxiliary maps. **Australian Journal of Soil Research**, v.41, p.1403–1422, 2003.

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

IBGE. **Manual técnico de pedologia**. 2.ed. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística, 2007. 300p.

JAFARI, A. et al. Spatial prediction of USDA—great soil groups in the arid Zarand region, Iran: comparing logistic regression approaches to predict diagnostic horizons and soil types. **European Journal of Soil Science**, Oxford, v.63, n.2, p.284–298, 2012.

JARVIS, A.et al. **Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90 m**. 2008.

KÄMPF, N.; GIASSON, E.; STRECK, E. V. **Levantamento pedológico e análise qualitativa do potencial de uso dos solos para o descarte de dejetos suínos da bacia do Rio Santo Cristo**. Porto Alegre: SEMA/RS, PNMA II, 2004.(Relatório final)

KLAMT, E. et al.**Solos do município de Dois Irmãos, RS:** Características, distribuição geográfica e aptidão de uso. Porto Alegre: Departamento de Solos, Faculdade de Agronomia, Universidade Federal do Rio Grande do Sul, 1993. 147p.

LACOSTE, M.; LEMERCIER, B.. WALTER, C. Regional mapping of soil parent material by machine learning based on point data. **Geomorphology**, Amsterdam, v.133, n.1–2, p.90–99, 2011.

LAGACHERIE, P.; HOLMES, S. Addressing geographical data errors in a classification tree soil unit prediction. **International Journal of Geographical Information Science**, London, v.11,p.183-198, 1997.

LAGACHERIE, P.; MCBRATNEY, A.B. Spatial soil information systems and spatial soil inference systems: perspectives for digital soil mapping. In: LAGACHERIE, P.; MCBRATNEY, A.B.; VOLTZ, M. (Ed.). **Digital soil mapping:** an introductory perspective. Amsterdam: Elsevier, 2007.p.3-24

LEMERCIER, B. Extrapolation at regional scale of local soil knowledge using boosted classification trees: a two-step approach. **Geoderma**, Amsterdam, v.171–172, p.75–84, 2012.

LOH, W-Y. Classification and Regression Trees. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,** Hoboken, v.1, p.14-23, 2011.

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

MENDONÇA-SANTOS, M.L.; DOS SANTOS H.G. The state of the art of brazilian soil mapping and prospects for digital soil mapping. In: LAGACHERIE, P.; McBRATNEY, A.B.; VOLTZ, M. (Ed.). **Digital Soil Mapping:** an introductory perspective. Developments in soil science. Amsterdam: Elsevier, 2006.p.39-54

MEYER, D.J. et al. Summary of the validation of the second version of the ASTER GDEM. **International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences**, Amsterdam, v. XXXIX-B4, p.291-293, 2012.

MINASNY, B., McBRATNEY, A.B. Incorporating taxonomic distance into spatial prediction and digital mapping of soil classes. **Geoderma**, Amsterdam, v.142, p.285–293, 2007.

MIRANDA, E.E. **Brasil em Relevo**.Campinas:Embrapa Monitoramento por Satélite, 2005.Disponível em: <http://www.relevobr.cnpm.embrapa.br>. Acesso em: 3 jun. 2014.

MOISEN, G. G. Classification and regression trees. In: JØRGENSEN, S. E; FATH, B. D. (Ed.). **Encyclopedia of Ecology**. Oxford :Elsevier, 2008.v.1, p.582-588

MOORE, I.D. et al. Soil attribute prediction using terrain analysis. **Soil Science Society of America Journal**, Madison, v.57, p.443–452, 1993.

MORAN, C.J.; BUI, E.N. Spatial data mining for enhanced soil map modelling. **International Journal of Geographical Information Science**. London, v.16,n.6,p.533-549, 2002.

NARDY, A. J. R.; MACHADO, F. B.; OLIVEIRA, M. F. A. As rochas vulcânicas mesozóicas ácidas da Bacia do Paraná: litoestratigrafa e considerações geoquímico estratigráfcas. **Revista Brasileira de Geociências**, São Paulo, v.38, p.178-195, 2008.

NELSON, M.; ODEH, I. Digital soil class mapping using legacy soil profile data: a comparison of a genetic algorithm and classification tree approach. **Australian Journal of Soil Research**, Melbourne, v.47, n.6, p.632-649, 2009.

OMUTO, C.T.; NACHTERGAELE, F.; VARGAS-ROJAS, R. **State of the Art Report on Global and Regional Soil Information: Where are we? Where to go?** Roma: Food and Agriculture Organization of the United Nations (FAO), 2013.69p.

QUINLAN, J. R. Induction of decision trees. **Machine Learning**, Dordrecht, v.1,n.1,p.81-106, 1986.

QUINLAN, J.R. **C4.5:**Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann, 1993. 20p.

RODRIGUES, T.L.; DEBIASI, P.; SOUZA, R.F. Avaliação da adequação dos produtos ASTER GDEM no auxílio ao mapeamento sistemático brasileiro. In: SIMPÓSIO BRASILEIRO DE CIÊNCIAS GEODÉSICAS E TECNOLOGIAS DA GEOINFORMAÇÃO, 3.,2010, Pernambuco-RE. **Anais**, 2010. 1CD-ROM

ROECKER, S. M. et al. A Qualitative Comparison of Conventional Soil Survey and Digital Soil Mapping Approaches. In: BOETTINGER, J.L. et al. (Ed.), **Digital Soil Mapping:**Bridging Research, Environmental Application, and Operation.Dordrecht: Springer, 2010. p.369-384.

ROKACH, L.; MAIMON, O. Z. **Data mining with decision trees:** theory and applications. Londres: World Scientific, 2008, 244p.

ROSSITER, D. G. **Technical Note:** Statistical methods for accuracy

assessment of classified thematic maps. Enschede (NL): International Institute for Geo-information Science & Earth Observation (ITC), 2004. 107p.

SANTOS, H.G. et al.**Sistema Brasileiro de Classificação de Solos**. 3 ed. rev. ampl. Brasília, DF, Embrapa, 2013. 353p.

SCULL, P.et al. Predictive soil mapping: a review. **Progress in Physical Geography**, London, v.27,n.2,p.171–197, 2003.

SCULL, P.; FRANKLIN, J.; CHADWICK, O. A. The application of classification tree analysis to soil type prediction in a desert landscape. **Ecological Modeling**, Amsterdam, v.181, n.1,p.1–15, 2005.

SKIDMORE, A. K. et al. An operational GIS expert system for mapping forest soils. **Photogrammetric Engineering & Remote Sensing**, Bethesda, v.62,p.501–511, 1996.

SMITH, M.P. et al. The effects of DEM resolution and neighborhood size on digital soil survey. **Geoderma**, Amsterdam, v.137, p. 58–69, 2006.

STEHMAN, S.V. Sampling Designs for Assessing Map Accuracy. In: INTERNATIONAL SYMPOSIUM ON SPATIAL ACCUARCY ASSESSMENT IN NATURAL RESOURCES AND ENVIRONMENTAL SCIENCES, 8., China, 2008.**Accuracy in Geomatics**. China, 2008.v.2, p. 8-15.

STEYERBERG, E.W. Validation of Prediction Models. In: STEYERBERG, E.W. (Ed.). **Clinical Prediction Models.** New York: Springer, 2009. p. 299-311.

TAGIL, S.; JENNESS, J. GIS-based automated landform classification and topographic, landcover and geologic attributes of landforms around the YazorenPolje, Turquia. **Journal Application Science**, [Faisalabad, Pakistan],v.8, p.910-921, 2008.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**, Boston, EUA: Addison-Wesley Longman Publishing, 2005. 769p

TEN CATEN, A. et al.Regressões logísticas múltiplas: fatores que influenciam sua aplicação na predição de classes de solos. **Revista Brasileira de Ciência do Solo**, Viçosa, v.35, n.1, p.53-62, 2011.

TEN CATEN, A. et al.Mapeamento digital de classes de solos: características da abordagem brasileira. **Ciência Rural**, Santa Maria, v.42, n.11, p.1989-1997, 2013.

TEN CATEN, A.et al. Spatial resolution of a digital elevation model defined by the wavelet function. **Pesquisa Agropecuária Brasileira**, Brasília, v.47, p.449-457, 2012.

THOMPSON, J. A.; BELL, J. C.; BUTLER, C. A. Digital elevation model resolution: Effects on terrain attribute calculation and quantitative soil-landscape modeling. **Geoderma**, Amsterdam, v.100, p.67-89, 2001.

TIMOFEEV, R. **Classification and regression trees (CART) theory and**

applications. 2004.39 f. Tese (Doutorado) – Center of Applied Statistics and Economics, Humboldt University, Berlim, 2004.

VALERIANO, M.M.;ROSSETTI, D.F. Topodata: Brazilian full coverage refinement of SRTM data. **Applied Geography**, Oxford, v. 32, p. 300-309, 2012.

VASQUES, G.M.; GRUNWALD, S.; SICKMAN, J.O. Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra. **Geoderma**, Amsterdam,v. 146, p.14–25, 2008.

WILSON, J. P, Digital terrain modeling. **Geomorphology**, Amsterdam, v.137,p.107-121, 2012.

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

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining:** practical machine learning tools and techniques.3 ed. San Francisco: Morgan Kaufmann, 2011.629p.

YOHANNES, Y.; WEBB, P. **Classification and regression trees, CART:** a user manual for identifying indicators of vulnerability to famine and chronic food insecurity**.** Washington: International Food Policy Rescue Institute, 1999.50p.(Microcomputers in policy research**,**v.3)

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, 2001.

8. RESUMO BIOGRÁFICO

Rodrigo Teske nasceu em 10 de agosto de 1982 na cidade de Blumenau, Santa Catarina (SC), Brasil. De 1989 a 2000 completou os estudos primários e secundários no Colégio Estadual Santos Dumont, em Blumenau, SC. Em dezembro de 2007 lhe foi concedido o Grau de Bacharel em Agronomia pelo Centro de Ciências Agroveterinárias da Universidade do Estado de Santa Catarina (UDESC), Lages, SC. Em abril de 2010, com estudos sobre gênese, morfologia, mineralogia e classificação do solo, conquistou o título de Mestre em Ciência do Solo pelo Programa de Pós-Graduação Ciência do Solo da UDESC, Lages, SC. Desde agosto de 2010 é Doutorando no Programa de Pós-Graduação em Ciência do Solo da Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, com estudos direcionados para a avaliação e a comparação de técnicas empregadas no mapeamento digital de solo.