DISSERTAÇÃO

MAPEAMENTO DIGITAL DE SOLOS COM A UTILIZAÇÃO DA TÉCNICA DE MINERAÇÃO DE DADOS E FERRAMENTAS DE GEOPROCESSAMENTO

JONAS DE ASSIS CINQUINI

Campinas, SP

2015

INSTITUTO AGRONÔMICO
CURSO DE PÓS-GRADUAÇÃO EM AGRICULTURA TROPICAL E SUBTROPICAL

MAPEAMENTO DIGITAL DE SOLOS COM A UTILIZAÇÃO DA TÉCNICA DE MINERAÇÃO DE DADOS E FERRAMENTAS DE GEOPROCESSAMENTO

JONAS DE ASSIS CINQUINI

Orientador: Jener Fernando Leite de Moraes
Co-orientador: Ricardo Marques Coelho

Dissertação submetida como requisito parcial para obtenção do grau de **Mestre** em Agricultura Tropical e Subtropical, área de concentração em Gestão de Recursos Agroambientais

Campinas, SP

2015

Ficha elaborada pela bibliotecária do Núcleo de Informação e Documentação do Instituto Agronômico

C575m Cinquini, Jonas de Assis
Mapeamento digital de solos com a utilização da técnica de mineração de dados e ferramentas de geoprocessamento. /Jonas de Assis Cinquini. Campinas, 2015. 84 fls.

Orientador: Jener Fernando Leite de Moraes
Co-orientador: Ricardo Marques Coelho
Dissertação (Mestrado) Agricultura Tropical e Subtropical – Instituto Agronômico

1. Mapeamento Digital dos Solos 2. Sistema de Informações Geográficas 3. Weka 4. Árvore de decisão I. Moraes, Jener Fernando Leite de II. Coelho, Ricardo Marques III. Título

CDD. 631.4

DEDICATÓRIA

A minha família, graças a compreensão, incentivos e conselhos, principalmente meu avô Luis Marques de Assis (*in memoriam*), grande entusiasta da ciência e minha mãe, por nunca deixar minha cabeça abaixar.

Aos meus amigos, que me auxiliaram dentro e fora do meio acadêmico.

AGRADECIMENTOS

Agradeço ao meu orientador Dr. Jener Fernando Leite de Moraes pelo comprometimento na minha formação e confiança no meu trabalho, desde os tempos de minha graduação, portando-se como um amigo em toda a minha trajetória.

Ao meu co-orientador Dr. Ricardo Marques Coelho pelos ensinamentos na ciência do solo e opniões construtivas em minha dissertação.

Ao Dr. Hélio do Prado, coordenador do projeto Ambicana, pelo fornecimento dos pontos de amostragem de solo, que tornou possível a realização do meu trabalho.

A todas as pessoas do Laboratório de Geoprocessamento do Instituto Agronômico de Campinas: Elisabete Monteiro da Silva, Tânia Maria Nicoletti, João Paulo de Carvalho, Nícia Marcondes Zingra, Alfredo Armando Carlstrom Filho que sempre me axuliaram em tudo que eu precisei durante toda minha estadia no laboratório.

Ao Instituto Agronômicco de Campinas, a todos os profissionais da Pós-graduação do Instituto Agronômico de Campinas e a todos os professores do curso de Gestão de recursos Agroambientais pela imensa contribuição na minha formação.

Aos meus colegas de turma, pela amizade e companheirismo, que me ajudaram e me ensinaram bastante durante todo o cumprimento de créditos do mestrado.

A todas as pessoas que de alguma forma contribuíram para que esse trabalho se tornasse realidade.

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

LISTA DE EQUAÇÕES

Mapeamento digital de solos com a utilização da técnica de mineração de dados e ferramentas de geoprocessamento.

RESUMO

O objetivo dessa pesquisa foi elaborar um mapa pedológico da carta SF-22-Z-A-I-4 (Paraguaçu Paulista), baseado na análise integrada de variáveis morfométricas e geologia, e comparar os resultados com um levantamento pedológico detalhado, obtido por métodos tradicionais, bem como avaliar o modelo treinado em área diferente de onde foi testado. O produto final foi obtido por um modelo de predição, construído por algoritmos de classificação, nos quais utilizaram variáveis oriundas do Modelo Digital de Elevação (MDE), geologia e classes de solo da carta topográfica SF-22-Z-A-I-2 (Quatá). A área de estudo foi dividida em área de treinamento (delimitada pelo mapa pedológico pré-existente dentro da carta Quatá) e área de teste do modelo (limite da carta de Paraguaçu Paulista). Foi elaborado um banco de dados geográfico dentro do Sistema de Informações Geográficas (SIG) Ilwis 3.3, onde gerou-se o Modelo Digital de Elevação (MDE), e a partir desse foram extraídas as variáveis de declividade, curvatura plana, curvatura em perfil, orientação das vertentes e índice topográfico combinado. Essas variáveis aliadas a geologia e pedologia formaram a matriz de dados para geração do modelo de classificação no Weka 3.7. Foi avaliado o algoritmo de classificação por árvore de decisão J48 com balanceamento das classes igual a 0,5 e sem balanceamento da matriz de dados de treinamento. Os modelos foram testados na área da carta de Paraguaçu Paulista, resultando em dois mapas pedológicos digitais na mesma área, com balanceamento de 0,5 e sem balanceamento das classes. Para avaliar o modelo resultante, o mesmo foi cruzado com classes de solo coletadas em campo. Os modelos avaliados apresentaram uma acurácia global de 67,67% e 58,27% quando avaliados sem a técnica de balanceamento de classes e com balanceamento igual a 0,5 respectivamente. O resultado da validação dos modelos foi de 49,25% de acurácia geral para o mapa de classes obtido sem balanceamento das classes e 44,8% com balanceamento das classes. Esses baixos valores podem ser explicados devido a utilização de um mapa pedológico elaborado com propósito de cultivo de cana-de-açúcar, com isso, as classes de relevo da região não foram fielmente contempladas. A variável geologia apresentou maior contribuição no modelo de mapa digital de solos, seguido pelas variável de altitude e declividade. Conclui-se que o mapeamento digital de solos constitui-se em uma ferramenta importante para a geração de mapas pedológicos, principalmente na escala de reconhecimento,

e novos recursos metodológicos, como o emprego de autômatos celulares, poderão contribuir para o avanço do MDS.

Palavras-Chave: Pedologia, Sistema de Informações Geográficas; Weka; árvore de decisão.

Digital soil mapping with the utilization of the data mining techniques and geoprocessing tools.

ABSTRACT

The aim of this research was to build a pedological map, of the sheet SF-22-Z-A-I-4 (Paraguaçu Paulista), based on the integrated analysis of the geology and morphometric variables, and compare the results with a detailed pedological survey, obtained from traditional methods, as well as evaluate a training model into a different area where it was tested. The final product was obtained by a prediction model build from classifications algorithms, which used the variables originated by the Digital Elevation Model (DEM), geology and soil classes from the sheet SF-22-Z-A-I-2 (Quatá). The study area was divided into train area (bounded by the pre-existing pedological map Quata), and model test area (Paraguaçu Paulista's topo map). Has been built a geographic database inside the Ilwis 3.3 Geographic Information System (GIS), where has been created the DEM, and from the DEM has been extracted the slope, plan curvature, profile curvature, aspect and topographic wetness index variables. The morphometric variables, plus geology and pedology made up the data matrix to build the classification model into the Weka 3.7. Has been evaluated the decision trees algorithm J48, with class balance equal 0.5 and without class balance of the train matrix data. The models has been tested at the Paraguaçu Paulista´s sheet, resulting into two digital soil maps from the same area, one which 0.5 class balance and another without class balance. To evaluate the resulted model, it has been crossed with pedological data field. The models showed a global accuracy of 67.67% and 58.27% when evaluated without the class balance and 0.5 class balance respectively. The result of the models validation was 49.25% global accuracy for the map without class balance and 44.8% with the class balance. The lower values of the accuracy could be explained, by the utilization of a pedologic map build to sugar cane crop purpose, so the slope classes of the region was not completely covered. The variable geology has showed the best contribution to the digital soil model, followed by the elevation and slope. Has been concluded that the digital soil mapping is a important tool for the pedological maps building, mainly at the survey scale, and new motodological sources, as the cellular automata, will help for the progress of the DSM.

Key Words: Pedology; Geographic Information System; Weka; decision tree.

1 INTRODUÇÃO

A grande extensão territorial do Brasil exige trabalho extenso e muitas vezes, de alto custo financeiro para realização de levantamentos pedológicos em escalas adequadas às diferentes aplicações desses mapas. Tendo em vista que aproximadamente 30 % (FAOSTAT, 2013) das terras brasileiras são utilizadas pela agropecuária, os levantamentos de solo são fundamentais para ampliação do conhecimento sobre os solos do país e também para a adoção de práticas conservacionistas mais adequadas.

As principais perdas de solos por erosão hídrica no Brasil são oriundas do manejo inadequado do solo nas atividades agrícolas (BERTONI E LOMBARDI NETO, 1990). Práticas agrícolas inadequadas aliadas à ocupação desordenada da terra são fatores determinantes para o esgotamento do solo. A obtenção de informações do solo e suas características são nesse sentido, fundamentais para a adoção de práticas conservacionistas (SACHS et al., 2010).

O mapeamento tradicional de solos tem sido substituído por técnicas de mapeamentos digitais, devido principalmente aos extensos trabalhos de campo necessários para levantamento de solos e características da paisagem de toda a área, além das análises em laboratório e pesquisas, que refletem em elevados gastos de dinheiro e tempo. Enquanto que os mapas digitais são elaborados utilizando pequena amostra de dados de campo, informações digitais do terreno, e menor tempo, utilizado para pesquisa e geração dos modelos matemáticos para serem rodados nos computadores. O mapeamento digital se iniciou com o recente avanço das técnicas em sensoriamento remoto, geotecnologias aliadas aos recursos computacionais atuais. A aplicação da técnica de mineração de dados e de sensoriamento remoto, aliada a utilização de Sistemas de Informações Geográficas (SIG), tem possibilitado predizer propriedades do solo em escalas regionais e locais (LEVI, 2012; CATEN et al., 2011; MCBRATNEY et al., 2003).

A utilização de modelos de predição dos atributos e comportamento dos solos tem aumentado significantemente com o avanço da capacidade computacional em processar dados e suas técnicas de manipulação (JENNY, 1941; GESSLER et al., 1995). Desta maneira é necessário o aumento das pesquisas para aprimoramento das técnicas e métodos que utilizam a mineração de dados integradas aos SIG para popularização dos levantamentos de solos por todo território nacional, conservando sua qualidade.

A região oeste do Estado de São Paulo apresenta considerável área de utilização agrícola, caracterizada também por um significativo avanço da cultura da cana de açúcar, notadamente em áreas então ocupadas por pastagens. Tendo em vista que os levantamentos de solos são fundamentais

para o planejamento agrícola de novas áreas de exploração e considerando as dificuldades já mencionadas para realização de levantamentos pedológicos tradicionais, propõe-se no presente trabalho a aplicação de metodologia de mapeamento digital de solos, com base em associações de variáveis morfométricas, de geologia e levantamentos pedológicos existentes na área de estudo.

Nesse contexto o presente trabalho tem como objetivo a elaboração de um mapa pedológico baseado na análise integrada de variáveis morfométricas, de geologia e comparar os resultados com um levantamento pedológico detalhado, obtido por métodos tradicionais.

A hipótese postulada é que nas áreas com predominância de relevo plano a suave ondulado, a acurácia na predição de classes de solo é menor, devido a menor diferenciação das variáveis derivadas do relevo.

Os objetivos específicos são:

- Gerar matriz de dados com informações morfométricas oriundas do modelo digital de elevação, litologia e pedologia;
- Gerar modelo de predição de classes de solos utilizando dados de treinamento da carta Quatá e o testando na carta de Paraguaçu Paulista;
- Validar mapa digital de classes de solo com dados coletados em campo.

2 REVISÃO BIBLIOGRÁFICA

2.1 Levantamento de Solos

O solo, diferentemente das plantas e animais, existe em um universo contínuo multivariado, pois muitas propriedades variam em diferentes proporções de um ponto a outro com poucas descontinuidades. Portanto o solo não tem limites naturais e indivíduos discretos, com isso pedólogos são forçados a criar limites artificiais, indivíduos e classes por definições arbitrárias das propriedades dos solos (EDMONDS, 2008).

Para representar a distribuição geográfica dos solos como corpos naturais, são realizadas práticas de levantamentos pedológicos. Segundo IBGE (2007) essas compreendem o registro de observações, análises e interpretações dos aspectos físicos e morfológicos dos solos, visando sua caracterização e classificação. Com o objetivo de subdividir áreas heterogêneas em parcelas com solos mais homogêneos, que apresentam a menor variabilidade possível. Os produtos obtidos podem ser representados por mapas e/ou textos explicativos, que tem como finalidade disponibilizar informações de auxílio à práticas agrícolas, manejo e conservação do solo, planejamento territorial e urbano, etc.

O nível de detalhamento dos mapas elaborados a partir dos levantamentos é definido com base no objetivo do estudo. Estudos em escala menores (nível regional, nacional) apresentarão maior generalização dos aspectos pedológicos da área, enquanto que mapas ou boletins realizados em escalas maiores (nível local) deverão fornecer maior detalhamento da área.

No Brasil, os levantamentos exploratórios (escalas entre 1:750.000 a 1:2.500.000) são os mais realizados, como exemplo o Projeto Radambrasil, que operou entre 1970 e 1985, no qual foram realizados levantamentos diversos dos recursos naturais da região amazônica utilizando imagens obtidas por radar (IBGE, 2007). Em escala diferente, a Seção de Pedologia do Instituto Agronômico de Campinas (IAC) realizou a partir de 1975 um levantamento semi-detalhado (entre 1:25.000 a 1:100.000) dos solos do estado de São Paulo, muito utilizados atualmente para planejamento territorial de nível local (IBGE, 2007). Portanto levantamentos de solos com diferentes escalas proporcionarão objetos de estudos e enfoques diferentes.

Em produtos de levantamentos de solos são informadas diversas unidades de mapeamentos. Essas são delineadas pelas mudanças das propriedades da paisagem e do solo. As unidades contendo predominância de apenas uma classe de solos, são classificadas como simples, enquanto que combinações de duas ou mais classes de solos distintos são nominados associações. Solos distintos

que apresentam limites poucos nítidos e de difícil individualização caracterizam unidades de mapeamento complexa (IBGE 2007).

As unidades de referências para o mapeamento, podem ser um indivíduo de solo até classes de solos (agrupamento de perfis de solos homogêneos), bem como unidades de mapeamento, que abrange um conjunto de áreas de solos com relações e posições definidas na paisagem (IBGE, 2007).

Os mapas obtidos pelos levantamentos podem ser realizados pelo método tradicional ou utilizando técnicas computacionais. Os levantamentos de solos tradicionais são realizados com observações em campo, que podem ser pontuais (análise do solo em perfis ou tradagem), por meio de transectos na paisagem (aspecto pedológico no contínuo da paisagem), ou até aéreas (RANZANI, 1969). Enquanto que o mapeamento digital é realizado por modelos computacionais a partir de informações do ambiente.

2.2 Mapeamento digital de solos

A alta demanda de pesquisas de proteção e conservação do solo no Brasil necessita de métodos que viabilizem os levantamentos de solos. Nesse contexto o mapeamento digital torna-se uma alternativa que demanda menos tempo e custo se comparada ao método tradicional de mapeamento de solos (MCBRATNEY, 2003).

O mapeamento digital de solos pode ser definido como a criação de sistemas de informações espaciais do solo por modelos numéricos para inferência de variações espaciais e temporais dos tipos de solos, a partir de observações e conhecimentos relacionados à variáveis ambientais (LAGACHARIE & MCBRATNEY, 2007).

Esse método tornou-se viável devido ao avanço da computação e de tecnologia da informação na manipulação de dados pedológicos. Aliado a isso, na ciência do solo, o aumento do poder das ferramentas como os Sistemas de Informações Geográficas (SIG), GPS, sensoriamento remoto e fonte de dados disponibilizadas por Modelos Digitais de Elevação (MDE), favorecem novas tendências de levantamentos de solos (MCBRATNEY, 2003).

O surgimento dos mapeamentos digital de solos ocorreu devido ao avanço da pedometria, ramo da ciência do solo que estabelece relações matemáticas e estatísticas entre variáveis quantitativas do solo (BUI, 2007).

Segundo HENGL (2003) a pedometria envolve diferentes campos científicos, desde a geoestatística até microbiologia do solo. O domínio da pedometria tem modificado ao longo do tempo

desde sua fundação, na qual atualmente é melhor definida como um campo interdisciplinar que envolve estatística aplicada, ciência do solo e geoinformação (Figura 01).

Figura 01. Pedometria pode ser considerada como uma ciência interdisciplinar, onde ciência do solo, estatística aplicada e ciência da geoinformação se cruzam.
Fonte: Adaptado de HENGL (2003).

Os avanços em pedometria facilitaram a formulação de modelos que estimam a ocorrência do solo na paisagem. MCBRATNEY et al. (2000) propuseram um modelo de predição dos atributos do solo (Equação 01) baseado em um modelo anteriormente proposto por JENNY (1941) com a seguinte equação:

$$Sc = f\ (s,c,o,r,p,a,n) \qquad [01]$$

Onde Sc (classe do solo) é função do solo (s), clima (c), organismos (o), relevo (r), material de origem (p), tempo (a), e localização (n).

Apesar de toda a potencialidade da utilização de modelos de predição e mapeamentos digitais de solo, sua acurácia é inferior à dos mapas convencionais, principalmente devido à variação do solo no espaço e incertezas das variáveis ambientais preditoras utilizadas (CATEN, 2008). Por tratar-se de modelos matemáticos que representam a realidade, os mapeamentos digitais apresentarão sempre alguma inconsistência dos dados inferidos da paisagem.

No entanto os avanços em pesquisa e técnicas tendem a melhorar a acurácia dos levantamentos de solos, tendo em vista o maior treinamento das pessoas para trabalhar com esta nova forma de mapeamento de solo, bem como o avanço das técnicas de modelagem e informações disponíveis (MCBRATNEY, 2003).

2.3 Relação solo-paisagem

Juntamente com o avanço das pesquisas na quantificação e mapeamento dos solos presentes no globo, a modelagem da paisagem também aparece como um grande desafio para as pesquisas atuais.

Para PRADO (2007) a paisagem é um retrato da ação combinada dos fatores de formação do solo (relevo, organismos, material de origem e clima ao longo do tempo). (Figura 02)

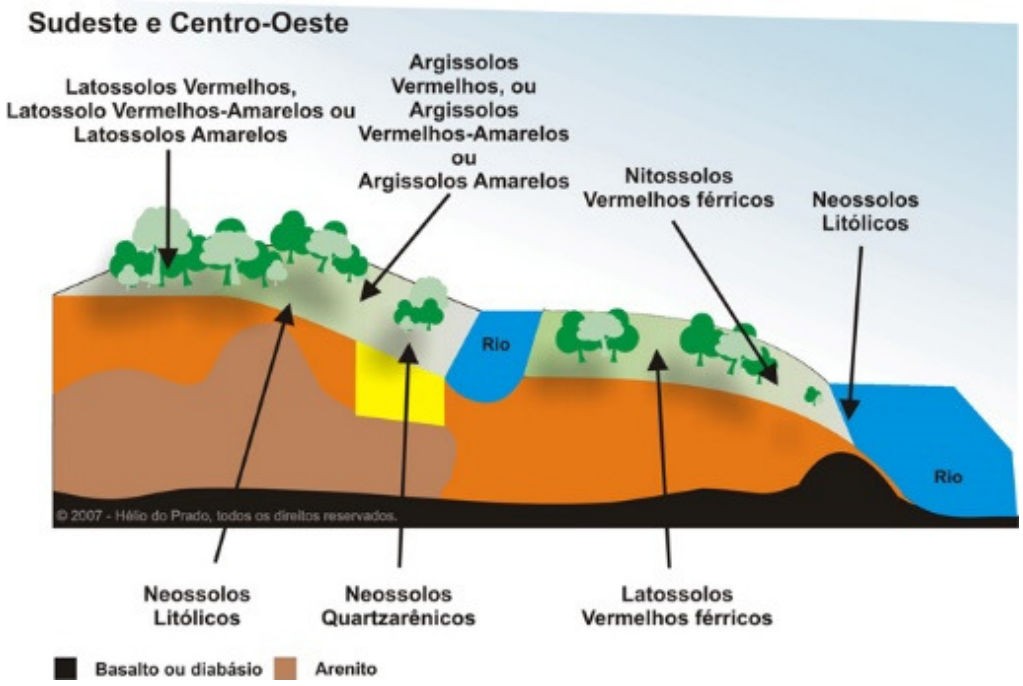


Figura 02. Relação solo-paisagem comum na região Sudeste e Centro-Oeste do Brasil.
Fonte: PRADO (2007).

Geralmente o relevo plano ou suavemente ondulado caracterizam paisagens mais estáveis, superfícies mais velhas, associadas a ocorrência de Latossolos. Enquanto que superfícies mais jovens são representadas por relevos mais ondulados ou forte ondulados, onde são encontrados Argissolos, Luvissolos, Cambissolos, Nitossolos, Chernossolos e Neossolos Litólicos. Nas áreas de baixadas planas há ocorrência de Vertissolos e nas várzeas de relevo plano os Organossolos e/ou Gleissolos.

O solo se comporta como um corpo contínuo na paisagem, consequentemente seu comportamento está intimamente ligado às características da mesma. O termo paisagem é geralmente utilizado para descrever a topografia (relevo) da região. Uma abordagem da relação dos solos com a paisagem são as topossequências, que explicam as diferenças do solo pela sua posição no relevo

(LEPSCH e BUOL 1974). Segundo COELHO et al. (1994) estudos das topossequências são bastante utilizadas no estabelecimento de relações entre atributos dos solos e da superfície.

Essa relação entre atributos geomorfológicos e o solo também é descrita por GESSLER et al (1995), quando dizem que as camadas de solo são resultantes da interação dos processos hidrológicos e a geomorfologia do terreno. Uma descrição da disposição, dimensão e natureza das camadas de solo, em locais na paisagem pode ser utilizado como uma ligação ou indicação para a distribuição espacial dos processos e vice-versa (GESSLER et al., 1995).

As relações entre solos e as formas de paisagem são, a muito tempo, a base para mapeamento de solos (IPPOLITI et al. 2005). SOUZA et al. (2004) destacam que a associação das formas da paisagem com a variabilidade espacial dos atributos do solo tem contribuído para trabalhos de levantamento e mapeamentos de áreas com classes de solos mais homogêneas.

A abordagem utilizada para representar a distribuição dos solos na superfície terrestre é a partir de modelos denominados solo-paisagem. Esses são fundamentais para o entendimento da relação espacial entre os atributos da paisagem e dos solos, com o objetivo de obter a compreensão da distribuição espacial dos atributos, características e comportamento dos solos (BROWN, 2005).

Os primeiros pesquisadores que construíram modelos relacionando solo com paisagem foram geólogos, ou seja, os mapas eram delineados de acordo com as formações geológicas, litologia e depósitos na superfície. Vale lembrar que esses modelos foram elaborados em escalas continentais. Enquanto que em 1935 foram propostos os primeiros modelos na escala de vertentes, no qual destaca-se o "catena", descrito por MILNE, G. (1936). O mesmo serviu como base para elaboração de modelos atuais e no qual procurou-se representar os processos de formação do solo pela caracterização de uma sequência solo-topográfica (BROWN, 2005).

Com enfoque para o planejamento ambiental modelos como o "catena" contribuíram para o desenvolvimento da modelagem solo-paisagem moderna, na qual são introduzidas técnicas de Sistemas de Informações Geográficas (SIG).

2.4 Técnica de mineração de dados

O avanço computacional a partir do século XXI tem permitido o processamento de grande quantidade de dados, em alta velocidade, e a execução de modelos de maior complexidade.

Nesse contexto os Sistemas de Informações Geográficas são fundamentais para simulações de fenômenos ambientais, incluindo distribuição espacial da ocorrência dos solos. Segundo ARONOFF

(1989) os SIG são conceituados como qualquer conjunto de procedimentos realizado de forma manual ou computacional com o objetivo de armazenar e manipular dados geograficamente referenciados.

A premissa dos modelos solo-paisagem na qual diz que, ao se determinar a relação entre cada solo e seu ambiente de ocorrência, pode-se inferir a localização de cada solo na paisagem pelas características do ambiente (HUDSON, 1992), pode ser aplicada utilizando SIG.

Técnicas de mineração de dados são fundamentais na exploração e manipulação de banco de dados georreferenciados, com o objetivo de encontrar padrões desses dados. Segundo HAN & KAMBER (2006) uma das tarefas na mineração é a classificação dos dados, na qual busca inferir uma variável dependente a partir de um conjunto de dados que contém atributos relacionados a essa variável.

Dentre as técnicas utilizadas para mapeamento digital de solos tem-se as técnicas de krigagem, que consistem em um método geoestatístico para interpolação espacial dos solos; lógica e os conjuntos nebulosos ("fuzzy"), método de classificação que permite alocação de indivíduos (pedons) em função de seu grau de pertinência a cada classe de solo; árvore de decisão, técnica de classificação por indução utilizadas para interações complexas entre atributos e classes de solos; redes neurais são modelos matemáticos capazes de trabalhar de forma similar ao cérebro humano (MENDONÇA-SANTOS e SANTOS, 2003). A técnica de árvore de decisão, foi escolhida para geração do modelo de classificação das classes de solo deste trabalho, devido aos resultados satisfatórios em estudos de média escala em relevo do interior paulista.

A indução por árvores de decisão é uma técnica de mineração de dados utilizada para classificação e predição de dados desconhecidos com base nos valores de atributos do conjunto de dados. Ou seja, com base em informações conhecidas cria-se um conjunto de dados de treinamento, do qual a árvore é montada e, a partir da mesma, há possibilidade de classificar os dados desconhecidos sem necessariamente testar todos os valores dos seus atributos (BREIMAN et al., 1984).

A árvore de decisão é um fluxograma com estrutura de uma árvore, onde cada nó interno (nó sem folha) denota um teste sobre um atributo, cada galho representa um resultado de um teste, e cada nó com folha (nó terminal) representa uma classe. O nó mais ao topo na árvore é o nó raiz (Figura 03).

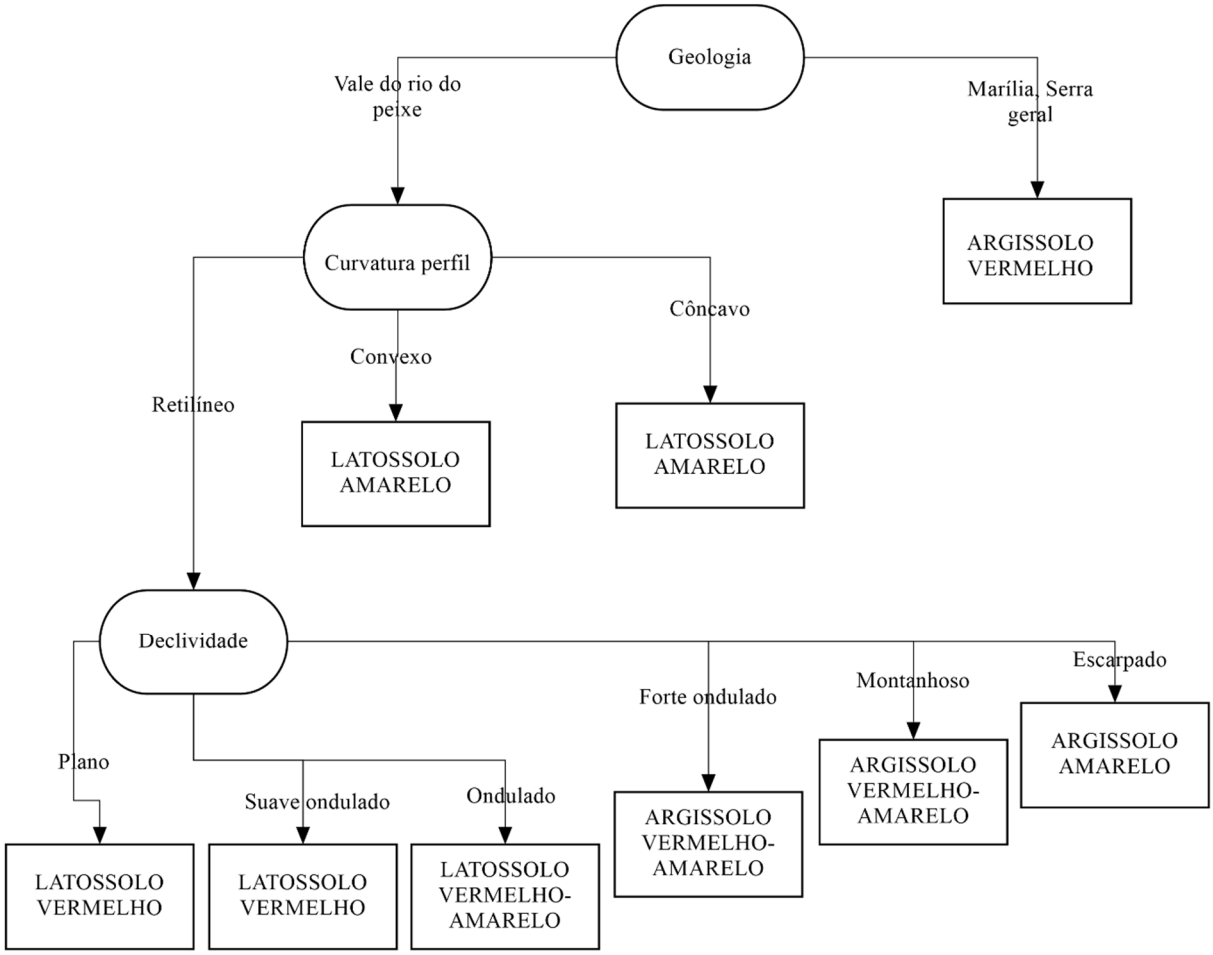


Figura 03. Exemplo de árvore de decisão para predição de classes de solo. Cada nó interno (seta) representa um teste de um atributo. Cada folha (retângulo) representa uma classe de solo predita. Fonte: Adaptado de HAN & KAMBER (2006).

O algoritmo que está oculto no método de indução por árvore de decisão é o chamado ID3 (*Iterative Dichotomiser*) desenvolvido por J. Ross Quinlan no final dos anos 1970 e início dos anos 1980 (AMO, 2004), que posteriormente após sua expansão se transformou no C4.5, com E. B. Hunt, J. Marin, e P. T. Stone em estudos sobre conceito de sistemas de aprendizagem (HAN & KAMBER, 2006).

Os algoritmos ID3 e C4.5 possuem uma abordagem na qual as árvores de decisão são construídas de cima para baixo (descendente) sem possibilidade de retorno para nós superiores e de maneira repetitiva com ramificação até o nó terminal. Ou seja, inicia-se com uma amostra de treinamento e suas classes associadas. O grupo de dados é repetitivamente particionado em pequenos grupos, como uma árvore que vai sendo construída, até que cada nó represente apenas uma classe (BREIMAN et al., 1984).

As ramificações ou partições dentro da árvore ocorrem quando um atributo se sobressai sobre os outros. O atributo escolhido para um determinado nó é aquele que possui o maior ganho de informação. A partir daí inicia-se um novo processo de partição dentro da árvore. No caso das árvores voltadas para classificação, o critério de partição muito utilizado é baseado na entropia (ONODA & EBECKEN, 2001).

Segundo HAN e KAMBER (2006) dentro da Teoria da Informação a entropia foi explicada por Claude Shannon como uma medição de informação faltante (*missing information*), ou quantificação da incerteza, ou seja, a entropia mede o nível de certeza que se tem sobre um evento. Enquanto que o ganho de informação mede a redução na entropia, ou na incerteza, ao selecionar um atributo para classificação.

Portanto uma árvore de decisão tem como objetivo minimizar a entropia e também possuir o menor número de nós, pois quanto menor o número de ramificações (complexidade da árvore) obtidas mais consistente será a classificação dos dados de treinamento.

Cada caminho percorrido da raiz de árvore até um dos nós terminais (folha), ou seja, de um determinado atributo até uma classe predita, pode ser deduzido como uma regra. O grupo de regras resultante pode ser simplificado para melhor compreensão do modelo por um usuário (ROKACH and MAIMON, 2014).

A utilização das árvores de decisão é bem ampla na descoberta de conhecimento em banco de dados, principalmente devido a sua simplicidade, transparência, por ser autoexplicativa e pessoas com pouco conhecimento em mineração de dados pode utilizá-la (ROKACH and MAIMON, 2014), além de sua interoperabilidade facilitada com Sistema de Informações Geográficas (SARMENTO, 2010).

No entanto, toda técnica de modelagem envolve erros e incertezas, nas quais devem estar bem determinadas e o usuário consciente das particularidades e limitações da técnica. Tradando-se dos algoritmos de árvore de decisão, deve-se trabalhar com valores discretos e a manipulação de grandes quantidades de dados podem gerar uma estrutura complexa, o que dificulta na manipulação e gestão dos dados.

Os resultados obtidos da árvore de decisão devem ser verificados utilizando dados que não tenham sido utilizados para o treinamento da mesma. Esse procedimento permite estimar como a árvore manipula os dados, podendo também estimar a proporção de erros e acertos ocorridos (BRAZDIL, 1999).

2.5 Precisão e acurácia de mapas

A precisão e acurácia do mapa são identificadas pela variação dos atributos de solos dentro das unidades de mapeamento e das classes de solo, sendo que a escala de mapeamento determinará quão preciso e exato deverá ser o mapa (SILVA, 2000).

Tratando-se da classificação por árvore de decisão, segundo ROKACH e MAIMON (2014) o desenvolvimento de indicadores eficientes para se avaliar a qualidade da árvore de decisão gerada ainda é um problema que não está resolvido por completo.

Para uma avaliação eficaz do desempenho de uma árvore de decisão necessário que a escolha das amostras que serão utilizadas tanto para treinamento (indução) como para o teste (avaliação) sejam escolhidas de maneira correta (BASGALUPP, 2010). O método de amostragem mais utilizado para grande quantidade de dados é o de re-amostragem, no qual geralmente é atribuído 2/3 (dois terços) dos dados para treinamento do modelo e 1/3 (um terço) para o teste.

Em mapeamentos de solos digitais é necessário verificar quão precisos os dados classificados são apresentados nos mapas. Geralmente isso é determinado pela elaboração de uma matriz relacionando todos os dados utilizados para classificação e os que foram classificados corretamente (avaliados pela comparação com dados de campo) (STORY & CONGALTON, 1986).

A Matriz de Confusão demonstra a qual classe cada dado pertence e também a qual classe ele foi classificado. Pela tabela será definida a medida de desempenho que será utilizada para avaliar o classificador. (Tabela 01).

Tabela 01. Exemplo de uma matriz de confusão de duas classes.

	Classe predita	
Classe real	Positiva	Negativa
Positiva	VP	FN
Negativa	FP	VN

Fonte: BASGALUPP, 2010.

Na Tabela 01 Verdadeiros Positivos (VP) são dados que pertencem a classe positiva e foram corretamente classificados como tal pelo classificador; Falsos Positivos (FP) são os dados que pertencem a classe negativa, porém classificados incorretamente como positivas pelo classificador;

Verdadeiros negativos (VN) são os dados que pertencem a classe negativa e foram corretamente classificados como negativos pelo classificador; Falsos Negativos (FN) são os dados que pertencem a classe positiva, contudo classificados como negativos pelo classificador.

A avalição dessas matrizes pode ser determinada pela divisão do número de registros classificados corretamente pelo número total de registros utilizados na classificação, ou dividir o número total de registros classificados corretamente pelo número total de registros utilizados como referência. Mas segundo CONGALTON (1991) recomenda-se utilizar o índice Kappa (Ka) para mediar acurácia de uma classificação temática.

O índice Kappa (Equação 02), foi proposto por Cohen em 1960 como uma medida de concordância, a proporção de dados em concordância em um determinado caso.

$$K = \frac{P_o - P_e}{1 - P_e} \quad [02]$$

Onde Po é a proporção de concordância dos dados observados, e Pe é a proporção de concordância esperada. Os valores que estão na diagonal principal da Matriz de Confusão refletem os dados em concordância com o modelo e o restante dos valores são as classes não preditas pelo modelo (COHEN, 1968).

Segundo SARMENTO (2010) a medida de concordância entre os dados estimados e os dados de referência podem variar de 0 (zero) a 1 (um), onde zero indica ausência de concordância e um a total concordância.

A possibilidade de avaliar a acurácia dos mapas digitais de solos é vantajosa frente ao mapeamento tradicional, pois o segundo não apresenta medidas quantitativas para avaliação de sua precisão, sendo possível elaborar mapas digitais de solos com precisão e acurácia considerável, comparados a levantamentos tradicionais.

3 MATERIAL E MÉTODOS

3.1 Caracterização da área de estudo

A área de estudo é a região do município de Quatá/SP, envolvente pelas cartas planialtimétricas do IBGE, escala 1:50.000 folhas: SF-22-Z-A-I-2 (Quatá); SF-22-Z-A-I-4 (Paraguaçu Paulista) entre as latitudes 22°00’ e 22°30’ Sul e longitudes 50°30’ e 50°45’ Oeste (Figura 04).

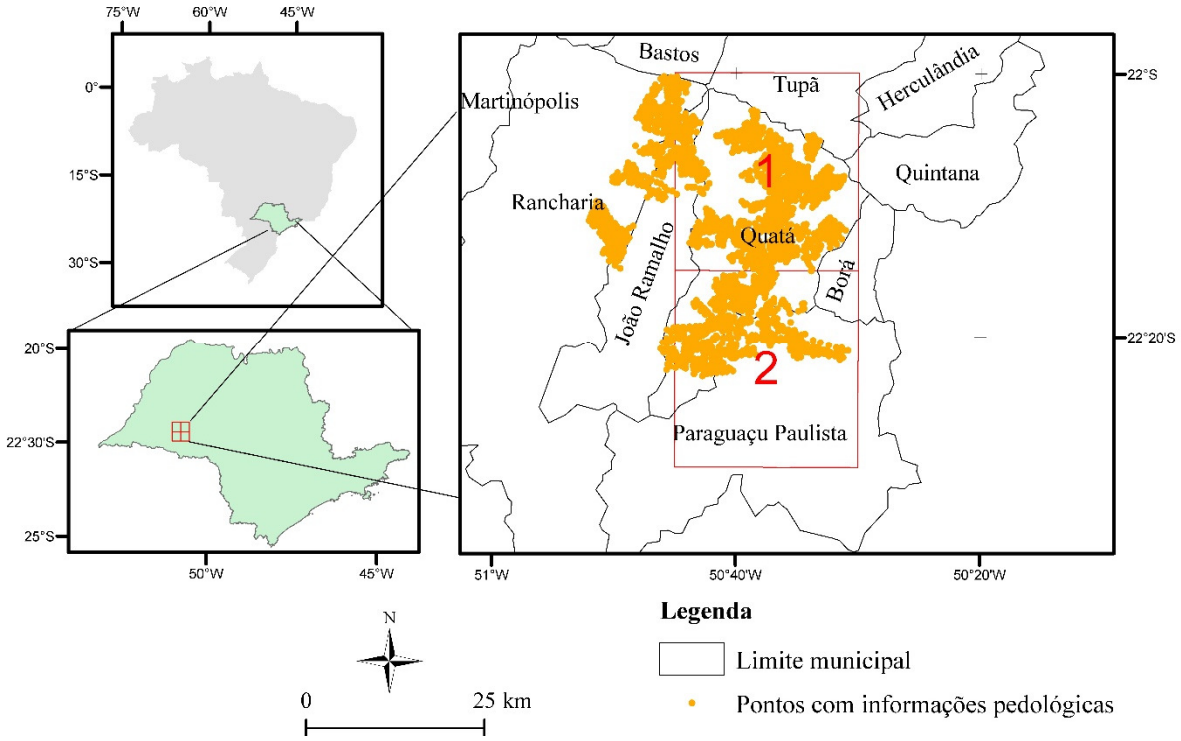


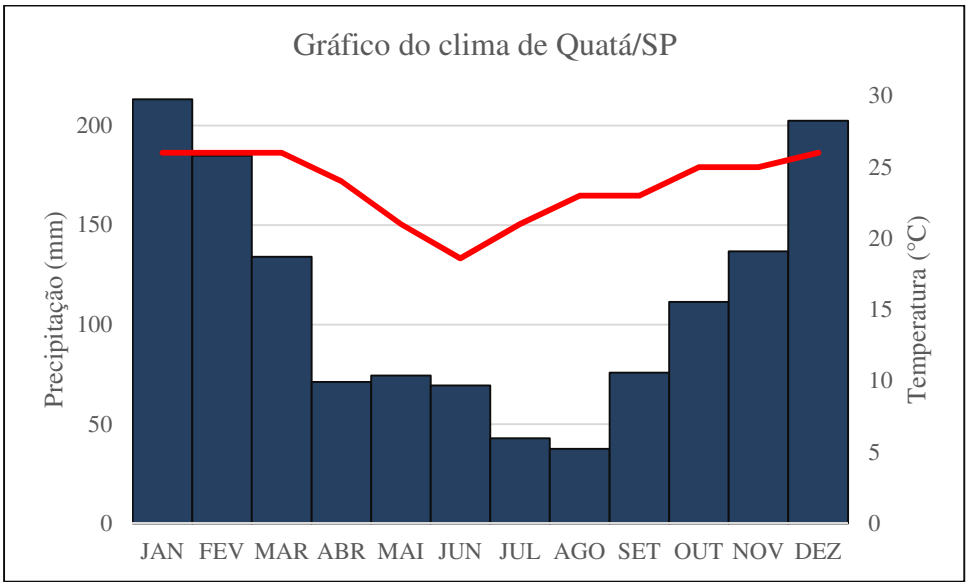
Figura 04. Localização geográfica da área de estudo.1 – Carta topográfica Quatá. 2- Carta topográfica de Paraguaçu Paulista.

Fonte: Elaborado pelo autor com dados de IBGE, 2013.

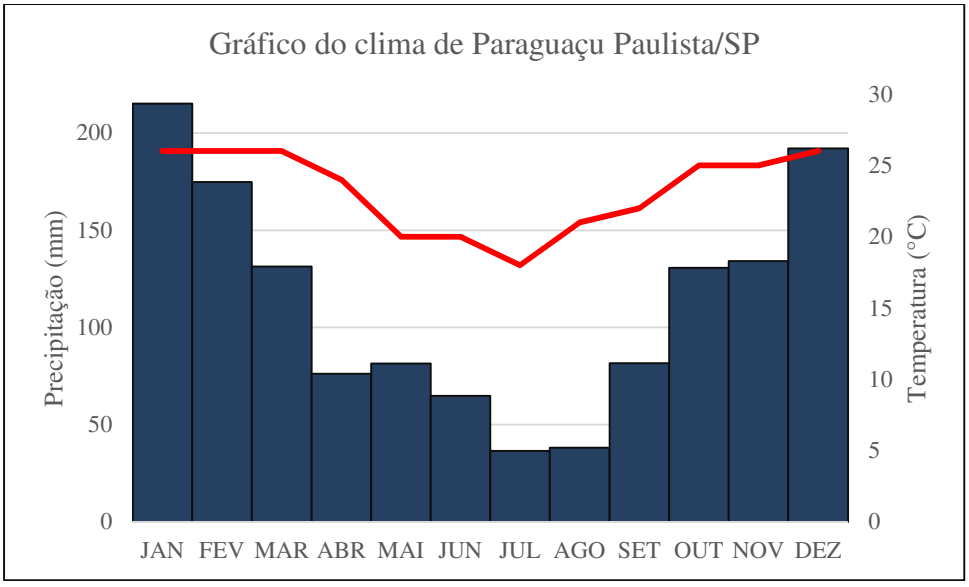
As cartas topográficas juntas possuem uma extensão de 142.624 ha, dos quais aproximadamente 85% nos municípios de Quatá e os 15% restante nos municípios de Bastos, Borá, Herculândia, João Ramalho, Quintana, Rancharia e Tupã.

3.1.1 Clima

O clima do município de Quatá (Figura 06a) é classificado como Aw, dentro da classificação climática de Koeppen, característico de regiões quentes situadas mais a oeste e noroeste do estado de São Paulo. Nomeado como tropical chuvoso com inverno seco e nos meses frios com temperaturas média superior a 18°C (CEPAGRI, 2015).



(a)



(b)

Figura 05. Clima dos municípios de Quatá e Paraguaçu Paulista.

Fonte: Elaborado pelo autor com dados de CEPAGRI, 2015.

Com clima semelhante, o município de Paraguaçu Paulista (Figura 06b), segundo a classificação climática de Koeppen, é do tipo Cwa, caracterizado pelo clima tropical de altitude, com seca no inverno e chuvas no verão. As temperaturas médias mais elevadas são superiores a 22°C.

3.1.2 Geologia

A área de estudo está inserida na geologia da Bacia do Paraná, nos grupos Bauru e São Bento. Mais de 85% da área é caracterizada pela formação geológica Vale do Rio do Peixe (anteriormente chamada de Adamantina), enquanto que cerca de 9% pela Formação Marília e e 6% pela Formação Serra Geral (Figura 06)

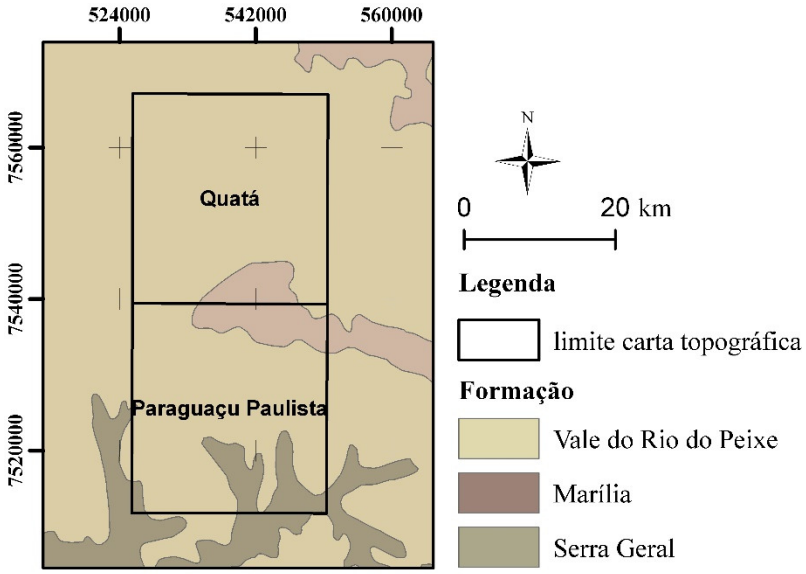


Figura 06. Formações geológicas presentes na área de estudo.
Fonte: Elaborado pelo autor com dados de CPRM (2006).

A Formação de maior expressão na área de estudo aflora em vasta extensão no oeste paulista, recobrindo unidades antigas do Grupo Bauru e também da Formação Serra Geral. Os sedimentos da Formação Vale do Rio do Peixe possuem granulometria fina, frequentemente compostos por mica. A mineralogia é predominantemente quartzo e a matriz composta por argilominerais bem distribuídos verticalmente (BONGIOVANNI, 2008). A Formação Marília, mais recente, recobre a Formação Vale do Rio do Peixe e possui arenito grosso, conglomerático, imaturo de coloração amarelo e vermelho (FERNANDES, 2004). Ambas pertencem ao período Cretáceo Superior mais recentes que a

Formação Serra Geral (Cretáceo Inferior) que compreendem um conjunto de derrames basálticos, intercalando camadas de arenito, litarenito e arenito vulcânico. Os derrames de basalto afloram na parte superior das *cuestas* basálticas e de morros testemunhos destas (BONGIOVANNI, 2008).

Segundo a relação solo-substrato geológico de BONGIOVANNI (1990) a Formação Serra Geral que apresenta litologia de rochas básicas apresenta classe de solo na ordem dos Latossolos Vermelhos e Nitossolos, a litologia de arenitos da Formação Vale do Rio do Peixe também apresenta formação de Latossolos, enquanto que os arenitos da Formação Marília estão associados a classe dos Argissolos.

3.1.3 Relevo

A litologia presente na área de estudo está inserida na geomorfologia do Planalto Ocidental (Almeida, 1964). Em sua maior parte, o relevo é caracterizado por espigões extensos com topos convexos ondulados, que compõem colinas extensas avançando no sentido dos tributários do rio Paraná.

Segundo IPT (1981) a Província Planalto Ocidental está dividida em quatro Zonas Geomorfológicas: Planalto de Marília, Planalto da Catanduva, Planalto de Monte Alto e Áreas Indivisas (Figura 07).

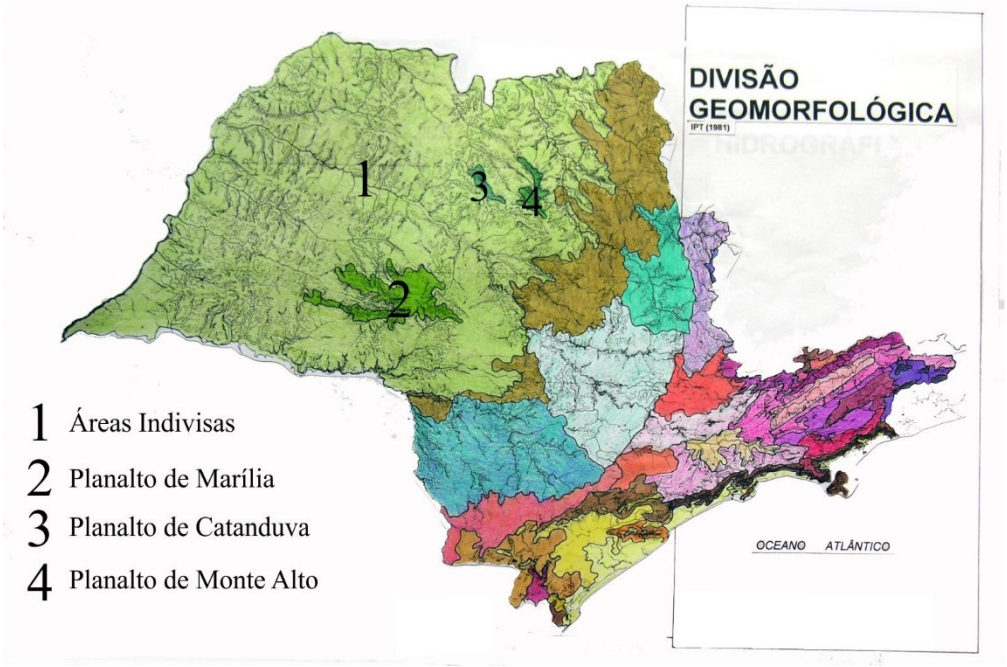


Figura 07. Esboço geomorfológico do estado de São Paulo.
Fonte: Adaptado de IPT (1981).

Outra abordagem é a de Ross e Moroz (1997), na qual a porção ocidental do Estado de São Paulo é denominada unidade morfoestrutural Bacia Sedimentar do Paraná e unidade morfoescultural Planalto Ocidental Paulista, ambas nomenclaturas representam áreas semelhantes da Província Planalto Ocidental descrito por ITP (1981).

Ambos autores identificam na região de Paraguaçu e Quatá um relevo levemente ondulado com presença de colinas amplas com topos aplainados e colinas médias com topos estreitos. No norte do município de Paraguaçu Paulista situam-se as colinas médias com altitude que varia de 500 a 610 metros, essas associadas a Formação Marília, enquanto que as Colinas Amplas presentes em relevo aplainado com altitudes entre 400 a 500 metros associam-se a Formação Vale do Rio do Peixe e Formação Serra Geral (BONGIOVANNI, 2008).

3.1.4 Pedologia

Os solos existentes na região do médio-Paranapanema, em sua maioria, caracterizados por terem alteração praticamente que total de seus minerais originais, condição essa influenciada pelo clima que atua sobre a região (subtropical, quente e úmido). Esses solos são representados pelas seguintes classes pedológicas: Latossolo Vermelho (LVd e LVdf), Nitossolo Vermelho (NV) e Argissolo Vermelho (PV) (CPTI, 1999). Em menor proporção também são encontrados solos pouco desenvolvidos, compostos ainda por quantidade expressiva de materiais da rocha mãe. São representados pelas classes Neossolo Litólico e Neossolo Quartzarênico.

Segundo a relação solo-substrato geológico elaborado por BONGIOVANNI (1990) na região entre Assis e Paraguaçu-Paulista, a formação geológica Serra Geral, dá origem a classe de solo Latossolo Vermelho Distrófico, Nitossolo Vermelho Eutroférrico e Latossolo Vermelho Distroférrico, enquanto que a Formação Adamantina (Vale do Rio do Peixe) origina as classes Latossolo Vermelho Distrófico, já na Formação Marília as classes de solos predominantes são Argissolo Vermelho-Amarelo.

CPTI (1999) apresenta características das unidades pedológicas na UGRHI do Médio-Paranapanema pela Tabela 02. Os autores descreveram as classes pedológicas inseridas na paisagem, isto é, contextualizando a ocorrência dos solos no meio físico (relevo e geologia).

Tabela 02. Relação solo, substrato geológico e relevo da Unidade de Gerenciamento dos Recursos Hídricos 17 (UGRHI-17) – Médio – Paranapanema.

Classe pedológica	Influência do substrato	Influência do relevo
Argissolo Vermelho-Amarelo (PVA)	Apresenta textura arenosa e média quando proveninente de arenitos (formações Adamantina, Marília e Santo Anastácio).	Desenvolve-se em relevos movimentados de colinas médias, morros e morrotes arredondados, etc.
Latossolo Vermelho (LVd)	Apresenta textura média quando proveniente de arenitos e textura arenosa quando originados de basaltos.	Desenvolve-se em relevos de colinas amplas. Quando existente em relevos mais movimentados constituídos por colinas médias, morros arredondados, é resultado de pedogênese sobre colúvios.
Latossolo Vermelho (LVdf)	Desenvolve-se a partir de rochas básicas (Formação Serra Geral)	Desenvolve-se em relevos de colinas amplas e topos aplainados.
Nitossolo Vermelho (NV)	Desenvolve-se a partir de rochas básicas (Formação Serra Geral)	Ocorre em relevos movimentados constituídos por colinas médias e morrotes alongados. Quando associado a Latossolo Vermelho desenvolve-se em enconstas mais declivosas nas proximidades de fundos de vales.
Neossolo Litólico (RL)	Ocorre em praticamente todas as formações geológicas, com texturas variadas dependendo da composição meniralógica do substrato.	Desenvolve-se em relevos muito movimentados.
Gleissolo (G)	Ocorre na associação a aluviões.	Presente em fundos de vales, várzeas e planícies aluviais.

Fonte: Adaptado de CPTI, 1999.

3.2 Método

3.2.1 Critério de seleção para área de estudo

A seleção da área de estudo foi realizada de acordo com os dados utilizados do projeto Ambicana (Ambientes de Produção para Cana de Açúcar), desenvolvido pelo Centro de Cana do Instituto Agronômico. Nessa área foram tomadas 1999 amostras (mil novecentos e noventa e nove pontos) de solo e obtidas as respectivas classificações pedológicas de cada uma de acordo com EMBRAPA, 2006. Posteriormente foram elaborados os respectivos mapas pedológicos das áreas de fornecedores de cana de açúcar. Dentro das metodologias atuais publicadas para predição de classes

pedológicas, são escassos os trbalhos que utilizem um extenso banco de dados para auxílio a elaboração de mapas digitais pedológicos utilizando atributos do relevo e geologia.

3.2.2 Base de dados

Foi gerado um banco de dados geográfico que contempla tanto a área de estudo de treinamento (carta Quatá) como a ser testada (carta de Paraguaçu Paulista).

Para manipulação dos dados foram utilizados os softwares computacionais ArcGIS 9.3 (ESRI, 2008), ILWIS (ITC, 2001), WEKA 3.7 (WEKA, 2014) e Microsoft Office Excel 2007.

Os dados de entrada utilizados foram divididos em matriciais (imagens, raster) e vetoriais (linhas, pontos, polígonos).

As imagens de entradas utilizadas foram as cartas topográficas de Quatá (folha SF-22-Z-A-I-2) e Paraguaçu Paulista (folha SF-22-Z-A-I-4) com informação de cotas altimétricas espaçadas a cada 20 metros, na escala 1:50.000, obtidos de IBGE (2013).

Os dados vetoriais de entrada utilizados foram os 1999 pontos de coleta de campo com informações de classes de solo e altitude distribuídos pela área de estudo, polígonos classificados com as classes pedológicas, polígonos com classes geológicas do estado de São Paulo obtido de CPRM (2006), e polígonos da divisão político-administrativa dos municípios próximo a área de estudo, obtidos de IBGE (2013).

Antes da manipulação dos dados foi elaborado um banco de dados geográficos inicialmente no Sistema de Informações Geográficas (SIG) ILWIS 3.3 e posteriormente no ArcGIS 9.3. Primeiro determinou-se um projeto delimitado por coordenadas geográficas no qual foram importados os dados, em seguida foi adotado um sistema coordenadas geográfica (*World Geodetic System* de1984 – WGS84) projetado para o sistema de coordenadas *Universal Transverse de Mercator* – UTM, Datum WGS84 Zona 22 do hemisfério Sul. Foi realizado esse procedimento para padronizar a base de dados, com isso torna-se possível cruzar diferentes planos de informações inseridos no banco de dados.

3.2.3 Processamento dos dados

Inicialmente foram delimitadas as áreas para treinamento e avaliação do modelo. Foram utilizadas as áreas das classes de solos, obtidas pelo projeto Ambicana, dentro da folha Quatá para treinamento do modelo, enquanto que para a validação, utilizou-se toda a extensão da carta Paraguaçú

Paulista. As escolhas foram determinadas pela grande representação das classes de solo da região em ambas as cartas topográficas. A carta Quatá foi utilizada para treinamento do modelo por apresentar maior quantidade dos pontos coletados (65%), e a carta de Paraguaçu Paulista foi utilizada para teste do modelo por conter ainda quantidade significante para validação do modelo (45%).

Após seleção das áreas de treinamento e validação do mapa pedológico digital, o tratamento dos dados teve início com a geração do Modelo Digital de Elevação (MDE) para ambas áreas. O mesmo foi obtido pela técnica de interpolação dos valores de altitude das curvas de nível dentro do SIG ILWIS 3.3.

Para isso as cartas topográficas foram georreferenciadas (adotadas coordenadas geográficas sobre as coordenadas cartesianas da imagem) e posteriormente as curvas de nível foram vetorizadas atribuindo-as o valor de sua respectiva cota. A interpolação das isolinhas de altitude foi executada segundo (ITC, 2001) criando uma imagem que representa um modelo digital do terreno, onde cada pixel possui uma localização geográfica e um valor de altitude.

O MDE foi recortado para a área que abrange as cartas topográficas de Quatá e Paraguaçu Paulista. Posteriormente os dados foram exportados do SIG ILWIS para serem manipulados em um banco de dados geográficos no ArcGIS 9.3.

A partir da imagem de elevação foram gerados os atributos morfométricos, variáveis essas que segundo a literatura (CRIVELENTI, 2009; CHAGAS, 2006; CATEN et al. 2012) são necessárias para incrementar o modelo para geração do mapeamento de classes pedológicas. As variárveis morfométricas geradas foram: declividade, direção de fluxo, curvaturas (perfil e plana) e Índice Topográfico Combinado.

A variável declividade (θ) é a medida da taxa de alteração da elevação no nível da superfície, medida como um ângulo em graus (0 a 90°) ou como uma percentagem (zero a cem). Calculada pela razão entre a distância vertical (a) e horizontal (b) da superfície (Figura 08).

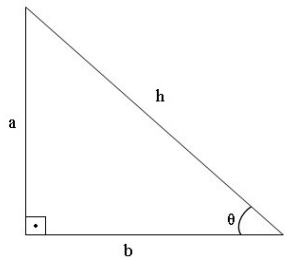


Figura 08. A declividade (θ), pode ser calculada a partir da distância vertical (a) e horizontal (b).

A derivada do modelo de altitude deve ser expressa em percentagem ou em graus. O grau de declive é expresso pelo valor do ângulo θ (Equação 03).

$$grau\ de\ declividade = \theta$$

$$\tan \theta = \frac{a}{b} \qquad [03]$$

A direção de fluxo (também conhecida como orientação das vertentes) é determinada pela direção da maior inclinação negativa do terreno. Expressa em graus (0° a 360°) ou direção geográfica (norte, nordeste, leste, sudeste, sul, sudoeste, oeste e noroeste). O algoritmo utilizado pelo ArcGIS 9.3 utiliza a Equação 04 para cálculo da direção de fluxo em cada pixel.

*Máxima queda = mudança no valor da altitude / distância * 100* [04]

As curvaturas são atributos topográficos baseados em segundas derivadas dos valores da matriz de elevação, calculadas pixel-a-pixel. A curvatura vertical é a direção para a declividade máxima, enquanto que a curvatura plana é perpendicular à direção da declividade máxima. A ferramenta de curvatura do software ArcGIS utiliza um algoritmo no qual a partir de uma janela de 3x3 pixels percorre toda a imagem com um polinômio de 4ª ordem (Equação 05).

$$Z = Ax^2y^2 + Bx^2y + Cxy^2 + Dx^2 + Ey^2 + Fxy + Gx + Hy + I \qquad [05]$$

O resultado da ferramenta de curvatura é a segunda derivada da superfície, ou também a declividade da declividade.

O ITC ou Índice Topográfico de Umidade (*Topographic Wetness Index – CTI*) foi desenvolvido por BEVEN e KIRBY (1979) que é definido segundo a Equação (06).

$$\ln \frac{\alpha_i}{\tan \beta_i} \qquad [06]$$

Onde α_i é a área de drenagem, de largura unitária e tang β_i é a é a declividade na superfície do ponto i.

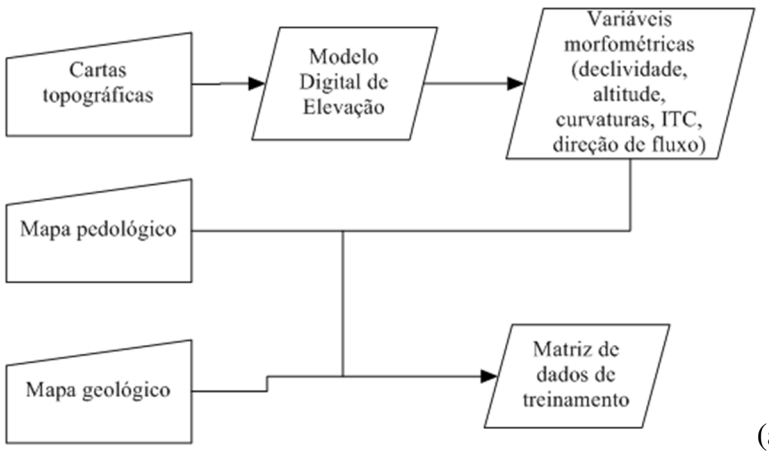
Todos os atributos morfométricos foram obtidos utilizando o conjunto de ferramentas do ArcGIS, apenas o ITC não possui uma ferramenta específica para o cálculo, por isso foi utilizado a Calculadora Raster sobre a imagem de elevação para realização dos cálculos.

As variáveis geomorfométricas, pedológicas e geológicas foram discretizadas para poderem ser cruzadas e inseridas no software de mineração:

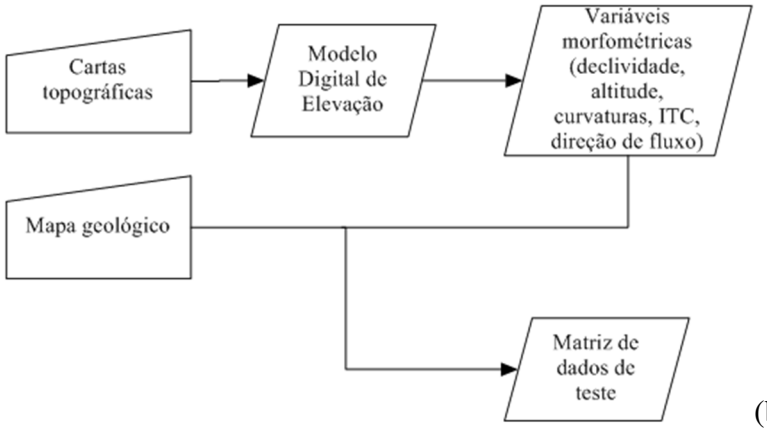
- Classes altitude (metros): 340 a 430, 430 a 450, 450 a 470, 470 a 490, 490 a 570 (determinadas pela técnica estatística Quartil);
- Classes geologia: Marília e Vale do Rio do Peixe (Adamantina);
- Classes de solos até o 2º nível categórico: Gleissolo Háplico, Latossolo Amarelo, Latossolo Vermelho, Latossolo Vermelho-Amarelo, Argissolo Amarelo, Argissolo Vermelho-Amarelo, Argissolo Vermelho, Neossolo Litólico, Neossolo Quartzarênico (EMBRAPA, 2006);
- Aspecto ou direção de fluxo: Norte, Nordeste, Leste, Sudeste, Sul, Sudoeste, Oeste, Noroeste;
- Curvatura plana: -1 a 0 (convergente), 0 (plana), 0 a 1 (divergente);
- Curvatura em perfil: -1 a 0 (convexo), 0 (plana), 0 a 1 (côncava);
- Declividade (%): 0 a 3 (Plano), 3 a 6 (Suave ondulado), 6 a 9 (Ondulado), 9 a 12 (Forte ondulado), 12 a 25 (Montanhoso), maior que 25 (Escarpado) (EMBRAPA, 2006);
- Índice Topográfico Combinado: 0-5, 5-10, 10-15, 15-20.

Cada variável foi representada por uma imagem contendo uma tabela de atributos representando as respectivas classes.

Para o treinamento do modelo foram utilizadas todas as variáveis discretizadas (Figura 09a), enquanto que para o teste do modelo, não foi utilizado o mapa de classes pedológicas (Figura 09b).



(a)



(b)

Figura 09. Fluxograma para obtenção das matrizes de treinamento (a) e teste (b).

O cruzamento foi realizado pela ferramenta "Combine" dentro do conjunto de ferramentas do ArcGIS. O mesmo procedimento foi realizado tanto para a matriz de treinamento como para o conjunto de dados testado pelo modelo.

As matrizes de dados foram padronizadas para serem importadas no software de mineração de dados (WEKA 3.7). Essa padronização constou dos seguintes procedimentos: retirada de inconsistências nas matrizes (valores faltantes, pixels com valores nulos ou extremos - ruídos), eliminação de informações fora da área de estudo e adequação ao formato dos dados para importação no WEKA.

A predição das classes de solo foi realizada por meio da técnica de classificação por árvores de decisão (HAN & KAMBER, 2006) com retirada de 30% dos dados da matriz para validação do modelo, e para geração do aprendizado de máquina foram utilizados 70% dos dados restantes.

Com o objetivo de não favorecer as classes com maior representatividade tornou-se necessário balancear as classes já determinadas. Pois ao buscar-se determinar a acurácia de cada classe (avaliada dividindo-se número de pixels atribuídos corretamente à classe pelo número total de pixels da categoria nos dados de referência), o resultado ficará favorável as classes com maior número de pixels (JENSEN, 1996). Os balanceamentos das classes foram 0 (zero) que representa a distribuição original dos dados e 0,5 (meio) que as classes são balanceadas de modo intermediário entre zero e um.

Em função da grande quantidade de dados utilizados na classificação, foi realizada a poda da árvore de decisão. Essa técnica tem como objetivo retirar pixels que não iriam contribuir com o modelo gerado (BATISTA, 2003). Segundo CRIVELENTI (2009) para mapas de solos na escala 1:50.000, podas com valores inferiores a 100 pixels não prejudicam o resultado final.

O resultado da árvore de decisão gerada é exportado do WEKA na forma de uma tabela, contendo a relação das classes preditas e sua localização geográfica. A tabela foi transformada em

raster que posteriormente é avaliado. O procedimento foi realizado para ambas áreas, área de treinamento (recorte carta Quatá) e a ser testada (recorte carta Paraguaçu Paulista).

Dentro do software de mineração a primeira operação para geração do modelo, foi a técnica da validação cruzada estratificada (*K-fold cross-validation*), realizada para cada base de dados (com e sem balanceamento). O método é utilizado para dividir a amostra de dados de treinamento em K conjuntos de testes dependentes, permitindo a construção de intervalos de confiança para a medida de desempenho. No caso, cada experimento foi repetido 10 vezes (K=10), para obtenção de estimativas mais confiáveis (JAIN & MAO, 2000 e WITTEN et al. 2011). Posteriormente os modelos foram avaliados com os dados de treinamento da carta Quatá e testados com a matriz de dados da carta de Paraguaçu.

A partir dos dois modelos gerados (classes balanceadas e não balanceadas) foram produzidos os mapas digitais. A avaliação de cada mapa gerado foi realizada com a criação da matriz de erro, onde foram confrontados os pixels da imagem gerada pelo modelo com pontos de solos coletados em campo na carta de Paraguaçu Paulista. Segundo CONGALTON (1991) uma matriz de erro é um quadro com um conjunto de dados dispostos em linhas e colunas que expressam um número de amostras de pixels assimiladas a uma categoria específica relacionados a uma classe verificada em campo.

A matriz de erros envolve técnicas de estatística descritiva e analítica para avaliação dos dados. O método mais simples é a acurácia global que é computada dividindo todos os valores que foram corretamente preditos (soma da diagonal principal) pelo total de número de pixels da matriz. Outras medidas de avaliação da classificação são pela acurácia do produtor e acurácia do usuário. A primeira é obtida pela somatória de todos os valores corretamente preditos de uma determinada categoria, dividido pelo total de pixels da mesma categoria. A exatidão do usuário é obtida pela divisão do número total de pixels corretamente classificados em uma categoria, pelo o número total de pixels que foi classificado pelo modelo na mesma categoria.

O coeficiente Kappa também foi calculado, estatística que mede a porcentagem dos valores da diagonal principal da matriz de erro e então ajusta esses valores para a quantidade de exatidões que poderia ser esperada para cada valor separado (WITTEN, 2011).

Os mapas foram elaborados pela exportação dos valores das classes preditos no Weka. Pela função "*Output predictions*" o software fornece uma lista com as seguintes informações: número da instância classificada (1–ARGISSOLO VERMELHO-AMARELO, 2-LATOSSOLO-VERMELHO, 3-ARGISSOLO AMARELO, 4-LATOSSOLO VERMELHO-AMARELO, 5-ARGISSOLO VERMELHO, 6-LATOSSOLO AMARELO, 7-NEOSSOLO QUARTZARENICO); classe atual de

solo (no caso do modelo teste as classes atuais eram desconhecidas - “?”); classe predita; erro na hora da classificação (classe atual versus classe predita); e acurácia da predição (Figura 10).

```
=== Predictions on test set ===

inst#,actual,predicted,error,prediction
1,1:?,2:LATOSSOLO_VERMELHO,,0.613
2,1:?,2:LATOSSOLO_VERMELHO,,0.613
3,1:?,2:LATOSSOLO_VERMELHO,,0.821
4,1:?,2:LATOSSOLO_VERMELHO,,0.821
5,1:?,2:LATOSSOLO_VERMELHO,,0.821
6,1:?,2:LATOSSOLO_VERMELHO,,0.821
7,1:?,2:LATOSSOLO_VERMELHO,,0.687
8,1:?,2:LATOSSOLO_VERMELHO,,0.687
9,1:?,5:ARGISSOLO_VERMELHO,,0.511
10,1:?,2:LATOSSOLO_VERMELHO,,0.653
11,1:?,2:LATOSSOLO_VERMELHO,,0.653
12,1:?,2:LATOSSOLO_VERMELHO,,0.653
13,1:?,2:LATOSSOLO_VERMELHO,,0.647
14,1:?,5:ARGISSOLO_VERMELHO,,0.781
15,1:?,2:LATOSSOLO_VERMELHO,,0.395
16,1:?,2:LATOSSOLO_VERMELHO,,0.395
17,1:?,2:LATOSSOLO_VERMELHO,,0.479
18,1:?,2:LATOSSOLO_VERMELHO,,0.479
19,1:?,2:LATOSSOLO_VERMELHO,,0.479
20,1:?,2:LATOSSOLO_VERMELHO,,0.479
21,1:?,2:LATOSSOLO_VERMELHO,,0.479
22,1:?,2:LATOSSOLO_VERMELHO,,0.479
23,1:?,2:LATOSSOLO_VERMELHO,,0.479
24,1:?,2:LATOSSOLO_VERMELHO,,0.479
```

Figura 10. Exemplo de saída do modelo de predição das classes de solo, extraído do Weka 3.7.

As classes preditas foram importadas para uma tabela e foram assimiladas as respectivas coordenadas geográficas de cada unidade de mapeamento. Dentro do SIG ArcGIS a tabela foi importada como uma feição de ponto, e esses foram transformados para um raster.

4 RESULTADOS E DISCUSSÃO

4.1 Produtos do MDE, geologia, variáveis morfométricas e pedologia

4.1.1 Área de treinamento do modelo de mapa digital de solos (recorte da carta topográfica de Quatá)

No treinamento do modelo de predição das classes de solo da carta Quatá foram utilizadas as variáveis morfométricas oriundas do modelo digital de elevação, informação litológica e classes pedológicas da carta topográfica de Quatá. A seguir essas estão apresentadas na seguinte ordem: MDE; geologia; orientação das vertentes; curvatura plana; curvatura em perfil; declividade, TWI e pedologia.

A partir das curvas de nível digitalizadas da carta topográfica de Quatá foi elaborado o mapa hipsométrico (Figura 11a), demonstrando que a classe de altitude com maior predominância situa-se entre 340 a 445 metros, localizada principalmente na porção norte da carta de Quatá, enquanto que a classe de 525 a 610 metros (maior elevação da área), ocorre na porção sul da área e correspondeu a 4,2% de todos os outros intervalos altimétricos.

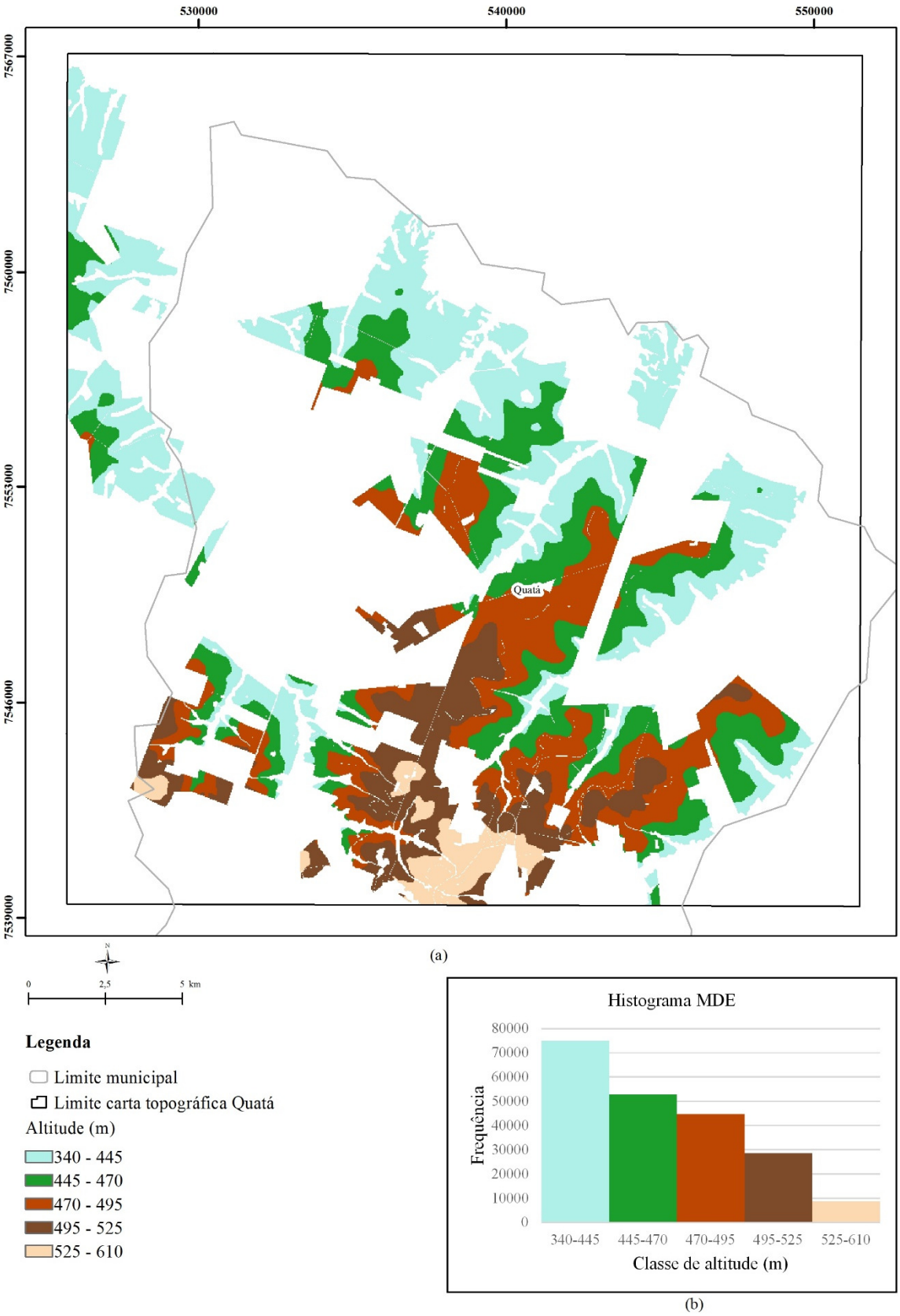


Figura 11. Mapa das classes de altitude (a) e histograma (b), para a área recortada na carta de Quatá.

Pela Figura 11b percebe-se, maior quantidade das classes de menores altitudes comparado as classes com maiores valores de elevação.

As classes de formação geológica presentes na área utilizada como treinamento do modelo são Marília e Vale do Rio do Peixe (Figura 12a). A primeira formação compõe-se de arenitos finos a grossos, imaturo, responsável pela sustentação do relevo mais pronunciado da região, onde está presente uma pequena cuesta, enquanto que a Formação Vale do Rio do Peixe, que corresponde a grande parte da outrora denominada como Formação Adamantina, é composta por arenitos muito finos, intercalados com siltitos ou limitos arenosos. Essa que representa quase que toda a área, com 83,74% de todo o recorte da carta topográfica (Figura 12b). A classe da formação Marília é encontrada em sua maioria na porção sul da carta de Quatá, e na fronteira entre as duas formações geológicas, percebe-se maior variedade das classes altimétricas, ou seja, diferenças no terreno, essas que influenciam também na classe e distribuição do solo presente. Nos locais onde há ocorrência da formação Vale do Rio do Peixe a classes de solo predominante é o Latossolo Vermelho, enquanto que na formação Marília presente em maiores altitudes apresentam classes pedológicas de Argissolos, Nitossolos, Neossolos e também Latossolos.

Figura 12. Mapa (a) e histograma (b) das formações geológicas para o recorte da área da carta Quatá.

A Figura 13a representa a orientação das vertentes. A presença da coloração mais clara no mapa denota o terreno voltado para a porção este (este, nordeste e sudeste), já as cores mais escuras representam a orientação das vertentes para o lado oeste (oeste, sudoeste e noroeste). A coloração

vermelha demonstra os locais voltados para o norte e a coloração cinza as faces planas. O histograma (Figura 13b) demonstra quantitativamente as classes das orientações do terreno, onde há predominância da classe este, com aproximadamente 31%, seguida das classes sudeste e oeste ambas com 11,6%. A classe noroeste aparece com menor expressividade (4,4%), juntamente com a classe plano (5,7%).

Como característica das orientações de vertentes presentes na área da carta Quatá, o terreno recebe energia proveniente do sol, podendo as áreas úmidas estarem presentes apenas no fundo dos vales próximos dos corpos d´água ou também na pequena proporção (7,13%) das vertentes voltadas para o sul.

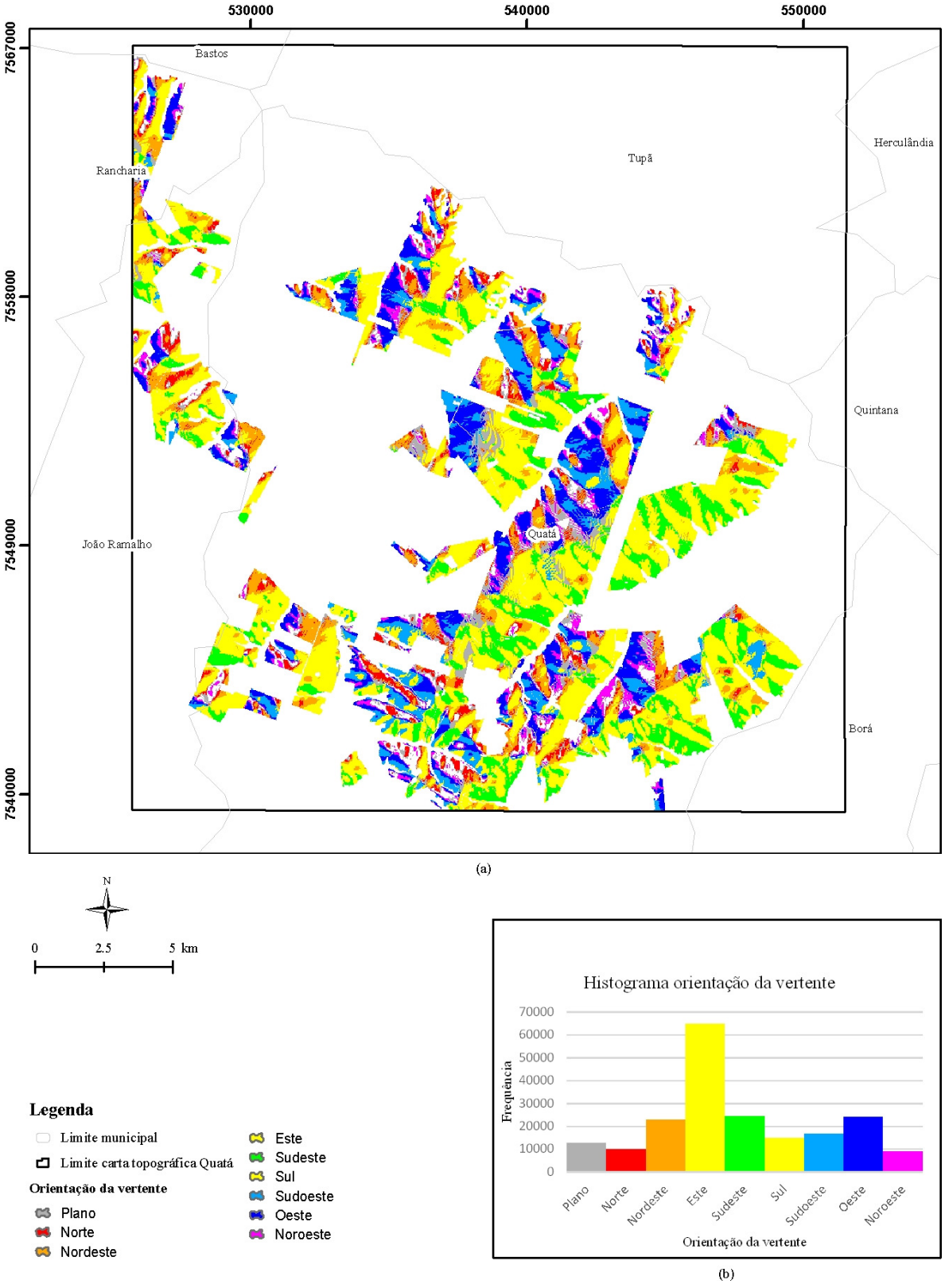


Figura 13. Mapa (a) e histograma (b) da orientação das vertentes para o recorte da área da carta Quatá.

Analisando o mapa de curvatura plana (Figura 14a) nota-se a semelhança nas classes convergente e divergente, destacando-se apenas a classe plano, principalmente na área central do

mapa. Quantitativamente (Figura 14b) não é diferente, a proporção de curvaturas planares que convergem e divergem são muito próximas (23,8 e 26,2% respectivamente), e os locais onde a curvatura horizontal é praticamente nula representam aproximadamente 50% da área total classificada na carta Quatá.

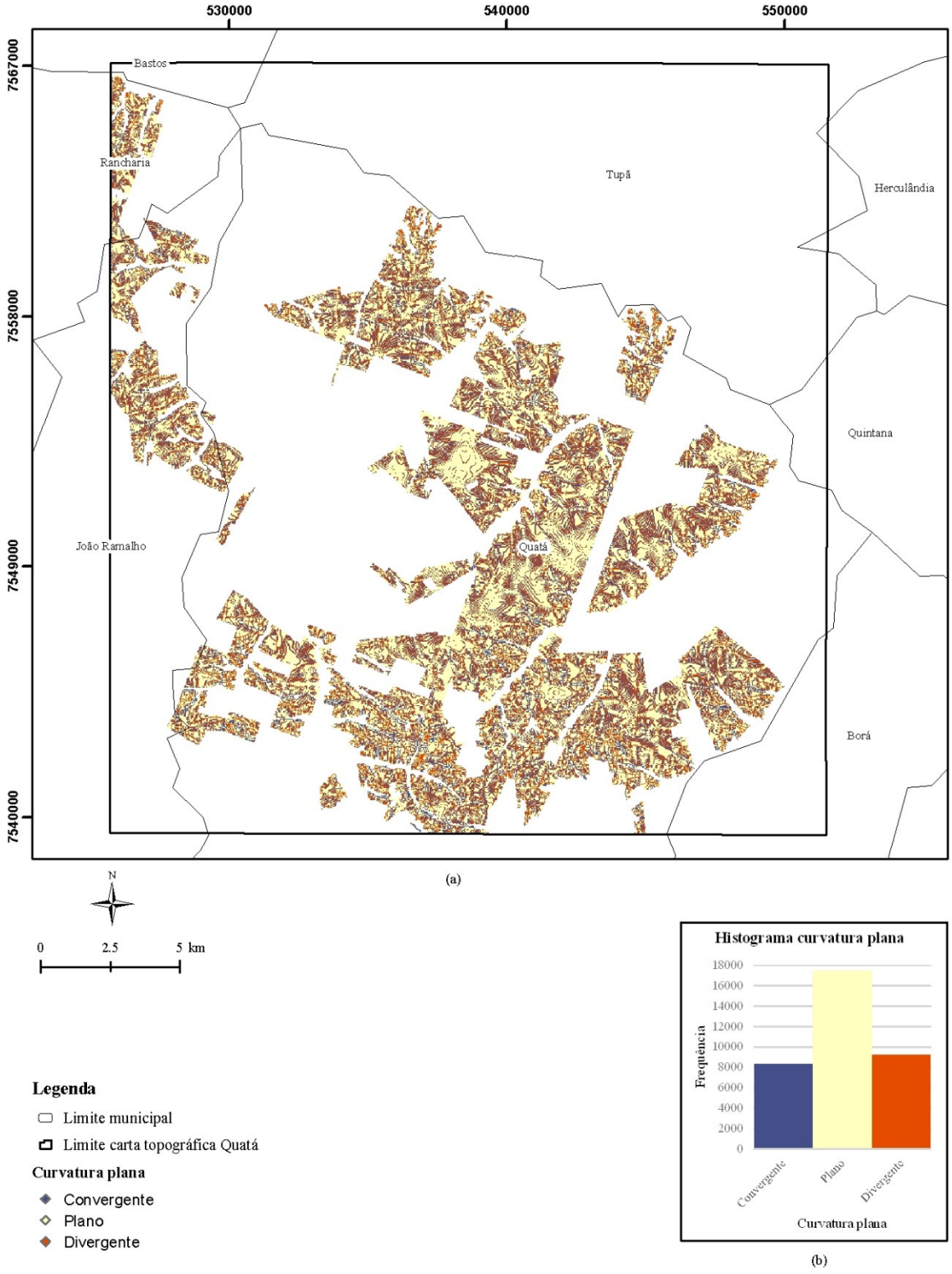


Figura 14. Mapa (a) e histograma (b) da variável curvatura plana da área recortada na carta Quatá.

Outro parâmetro derivado do MDE é a curvatura vertical, que é a vertente quando olhada em perfil. A Figura 15a representa espacialmente a curvatura em perfil e semelhante a curvatura horizontal, a coloração clara apresentou maior quantidade (dispersa por toda a área), que foi rotulada como a classe retilíneo (valores nulos da curvatura), quando comparada as outras classes. Enquanto que as curvaturas côncavas e convexas se confundem no mapa, pela semelhante proporção. De acordo com a Figura 15b, dentre as três classes da curvatura vertical, 50% é representada pelo perfil retilíneo, 26.2% pelo côncavo e 23,8% pelo convexo.

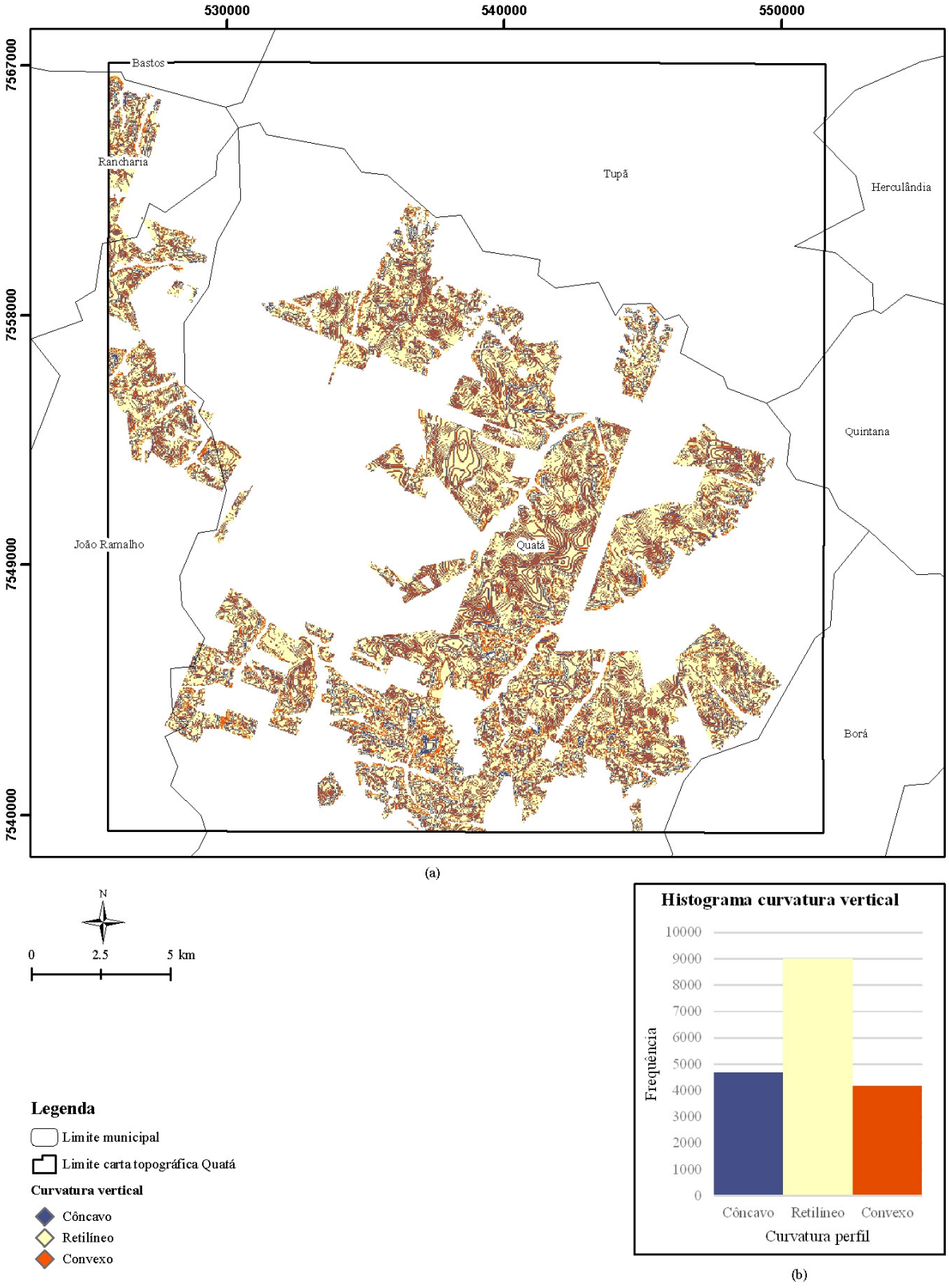


Figura 15. Mapa (a) e histograma (b) da variável curvatura em perfil.

Essa maior quantidade de valores da classe retilíneo demonstra a homogeneidade do terreno, quanto as variações da declividade, migração e acúmulo de matéria na superfície, processos de

pedogênese e morfogênese caracterizados por esse tipo de terreno com menor expressão da curvatura vertical. Outro fator relevante que refere-se à elevada proporção de valores de classes que representam a curvatura retilínea é a escala na qual o estudo está sendo realizado. Tendo em vista que as cartas topográficas foram elaboradas na escala de 1:50.000, que é considerada de semi-detalhe, as variações no terreno não são facilmente perceptíveis, ao contrário em estudos realizados em escalas maiores, com mais detalhes, onde provavelmente seria possível identificar feições mais detalhadas do relevo.

A partir do modelo digital de elevação também foi gerado o mapa de declividade (%) da área teste. O mapa das classes de declive (Figura 16a) demonstram as classes de relevo plano, suave-ondulado e ondulado com maior número de valores comparadas as outras classes. Na porção sul verificam-se um movimento mais acentuado no terreno predominando classes ondulado e forte ondulado. A classe de declive suave ondulado (declividade de 3-6% no terreno) representou 50% de toda a área utilizada para treinamento do modelo de mapa de solos, seguido pela classe plano e ondulado (23,6 e 20,8% respectivamente), e declives acima de 25% foram praticamente nulos na área (apenas 0,001% de todas as classes) (Figura 16b).

Figura 16. Mapa (a) histograma (b) das classes de declive da área de treinamento na carta Quatá.

A próxima variável morfométrica descrita indica a presença ou não de zonas de saturação de água, umidade em um determinado ponto (pixel, com área igual a 900 m^2) calculado a partir da característica do terreno, principalmente declividade e área de contribuição a montante. O mapa do

índice topográfico combinado (também conhecido como índice topográfico de umidade), (Figura 17a) demonstra maior frequência dos índices 0-25 e 0,25-0,5 comparado com o restante. Apenas pouco mais de 1% de todos os índices são representados por 0,5-0,75 e 0,75-1, enquanto quase 68% dos índices topográficos combinado é de 0-025 (Figura 17b).

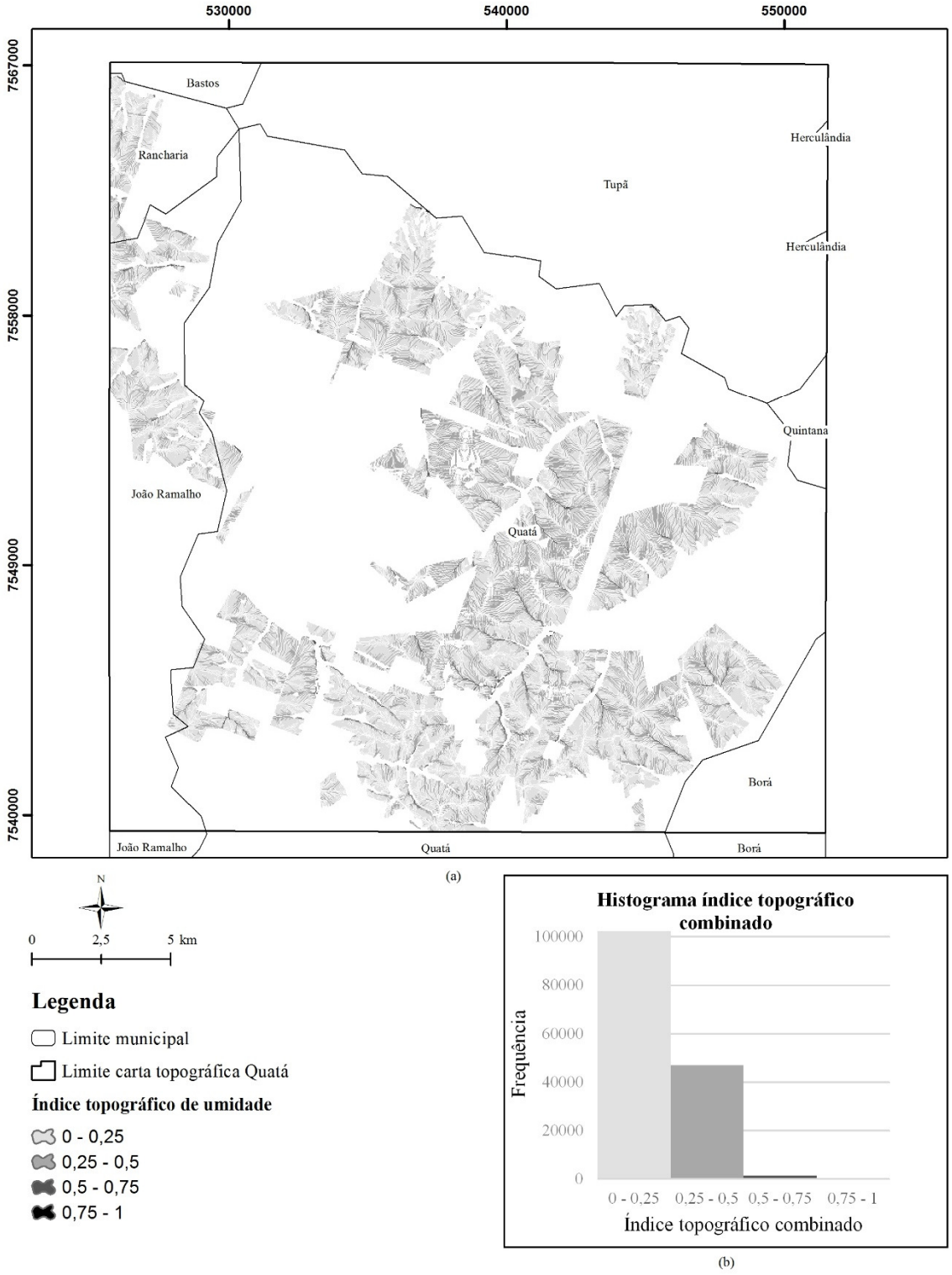


Figura 17. Mapa (a) e histograma (b) do índice topográfico combinado da área recortada dentro da carta Quatá.

O mapa de solos da carta de Quatá (Figura 18a) elaborado por intermédio da coleta pontual de amostras de solo foi utilizado juntamente com as variáveis anteriormente apresentadas para treinamento do modelo.

A distribuição espacial das classes de solos para a área de treinamento do modelo de mapeamento digital de solos apresenta grande proporção da classe LATOSSOLO VERMELHO na região central da carta Quatá. As classes aparecem mais diversificadas no sul da área, porém ainda com predominância de uma classe, nesse caso do ARGISSOLO VERMELHO. Pelo histograma da imagem (Figura 18b) verifica-se que as classes de solo ARGISSOLO AMARELO, GLEISSOLO HAPLICO, LATOSSOLO AMARELO, NEOSSOLO LITOLICO E NEOSSOLO QUARTZARENICO não atingem 1% do total das classes de solo cada uma, enquanto que apenas a classe LATOSSOLO VERMELHO representa 61,51% de todas as unidades de mapeamento. As classes LATOSSOLO VERMELHO-AMARELO, ARGISSOLO VERMELHO-AMARELO e ARGISSOLO VERMELHO apresentam juntas 36,4%.

Figura 18. Mapa (a) e histograma (b) das unidades de mapeamento de solo obtidas em campo recortadas dentro da carta Quatá.

4.1.2 Área teste do modelo de mapa digital de solos (carta topográfica de Paraguaçu Paulista)

Para testar o modelo de predição das classes de solo foram utilizadas as mesmas variáveis apresentadas na seção 4.1, com exceção da variável pedologia (que tem-se por objetivo predizê-la), porém na carta topográfica de Paraguaçu Paulista. A ordem de explanação das variáveis é a mesma do item anterior.

Ao sul da carta Paraguaçu Paulista a variação nos valores de altitude é mais acentuada, essa pode ser verificada pela presença da classe de altitude mais elevada (coloração branca, 525 – 610 metros) por toda a porção norte da carta Paraguaçu Paulista (Figura 19a). As classes de altitude predominantes são 470-495 e 495-525 metros (20,43 e 20,45% respectivamente). Outras classes com menor valor altimétrico, porém com frequência de ocorrência aproximadas foram as 340-445 e 445-470 metros, ambas com 19,7% de todas as classes.

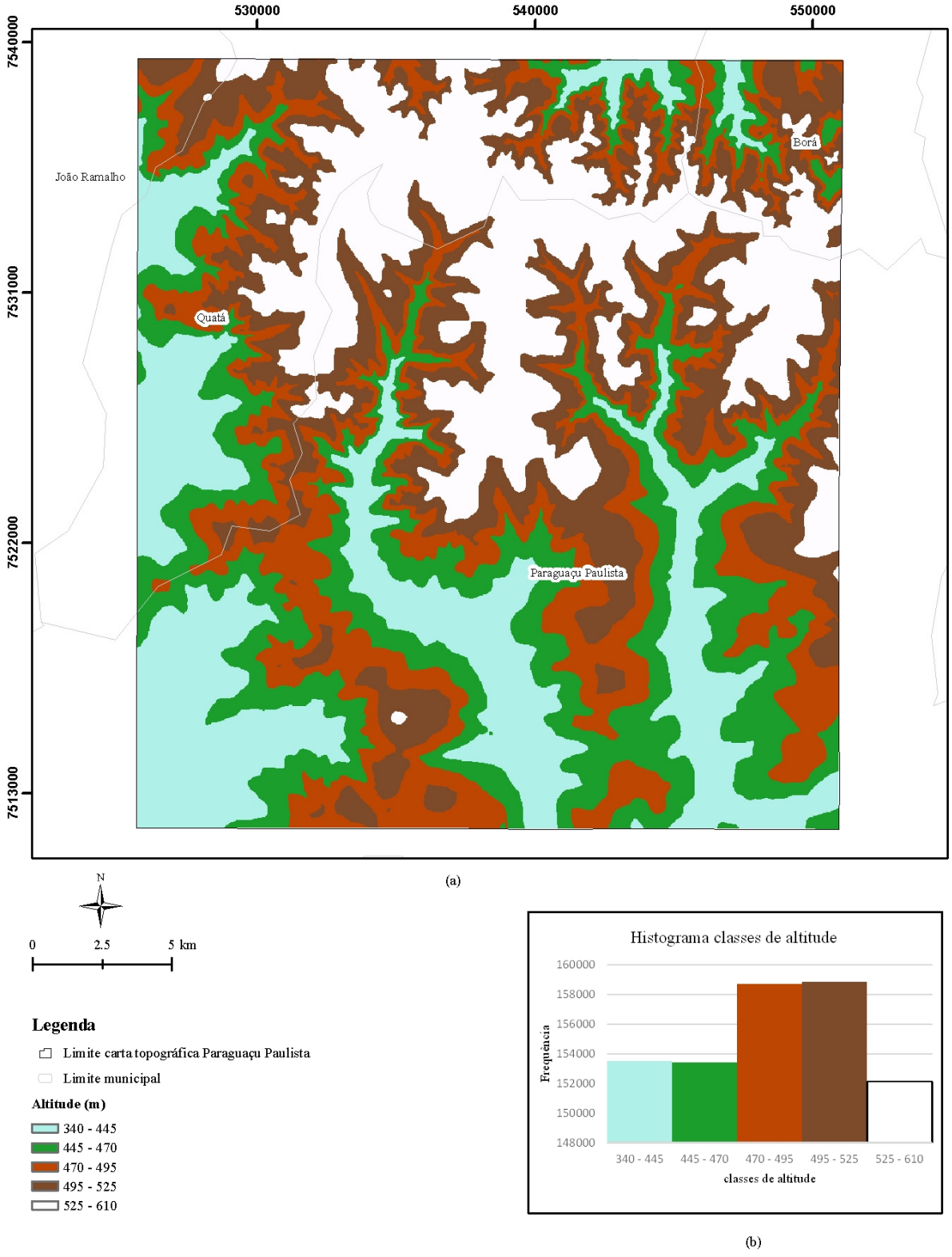


Figura 19. Mapa (a) e histograma (b) das classes de altitude presentes na carta Paraguaçu Paulista.

A formação geológica predominante na área da carta Paraguaçu Paulista é Vale do Rio do Peixe (anteriormente conhecida como Adamantina), pertencente ao grupo Bauru. No sul há ocorrência da formação Serra Geral, do grupo São Bento e uma porção ao norte da carta da formação

Marília também do grupo Bauru (Figura 20a). A formação Serra Geral constitui-se de basalto, com diferenças granulométricas, e intercalada com camadas de arenito.

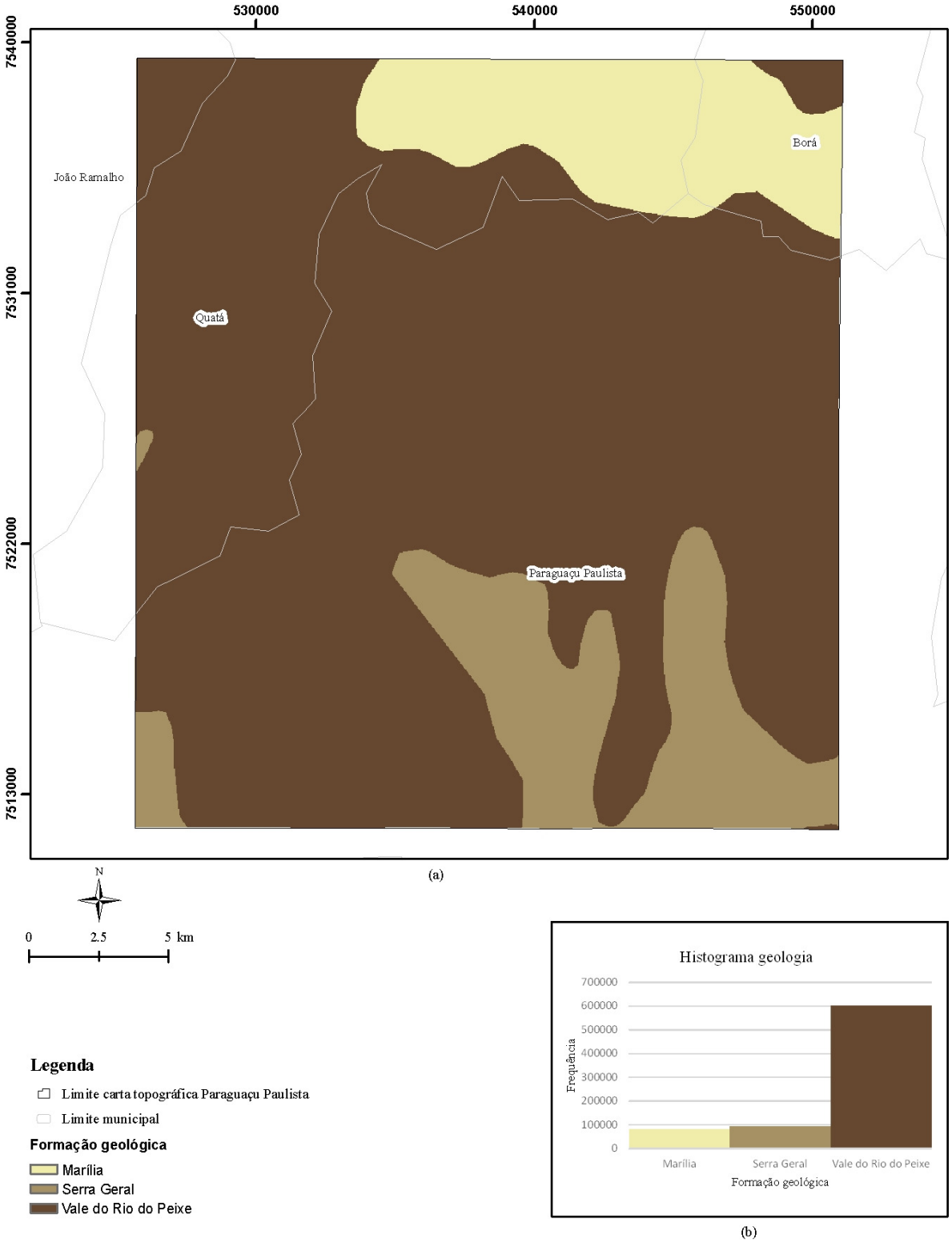


Figura 20. Mapa (a) e histograma (b) das formações geológicas presentes na carta Paraguaçu Paulista.

A formação que ocupa maior parte da área da carta apresenta aproximadamente 77% de toda a carta, enquanto que as formações menos expressivas recobrem 10.5 e 12% de toda a área (Marília e Serra Geral respectivamente).

A presença de uma nova feição geológica influencia na classe de solo presente no terreno, devido a diferença do material de origem da formação Serra Geral, que ao invés dos arenitos das formações Vale do Rio do Peixe e Marília, o basalto origina solos com texturas argilosas.

O mapa da orientação das vertentes (Figura 21a) demonstra predominância das classes Este, Sul, Sudoeste e Oeste. As faces dos morros voltadas para Este (cor amarela), localiza-se principalmente na porção centro-sul da carta, enquanto que as classes de orientação sul, sudoeste e oeste encontram-se distribuídas por toda a extensão da carta.

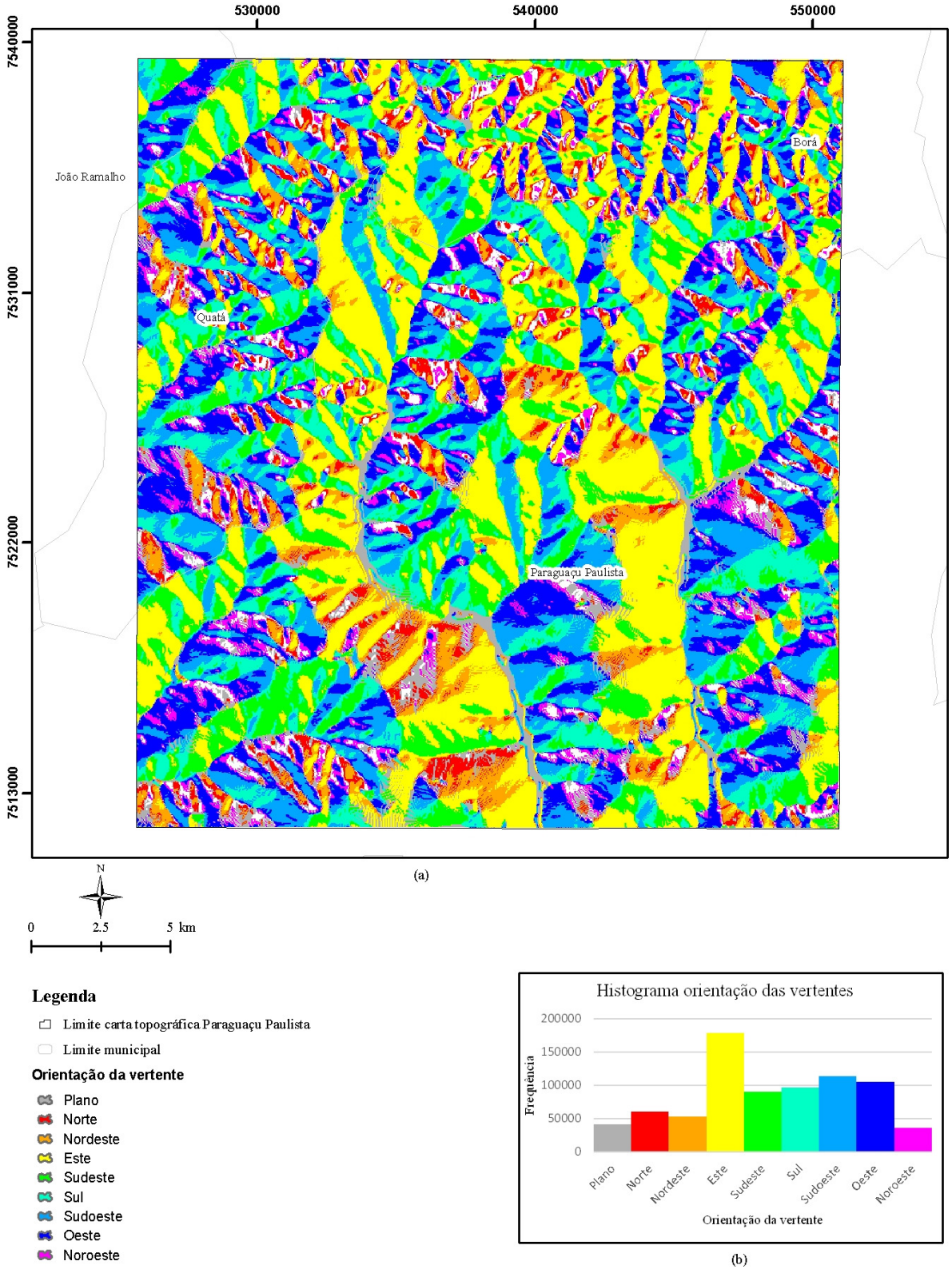


Figura 21. Mapa (a) e histograma (b) da orientação das vertentes na carta Paraguaçu Paulista.

A orientação da classe noroeste possui menor número de ocorrências por toda a área, com menor frequência também, as classes plano, nordeste e norte somam 20% do valor total de pixels na

imagem. As orientações sul, sudoeste e oeste representam juntas 40,6% e as classes este e sudeste representam 34,6% de todas as observações (Figura 21b).

No mapa de curvatura horizontal (ou plana), a classe plano apresentou maior frequência de ocorrência por toda a carta topográfica de Paraguaçu Paulista (Figura 22a), enquanto que as classes convergente e divergente apresentaram menor quantidade de pixels.

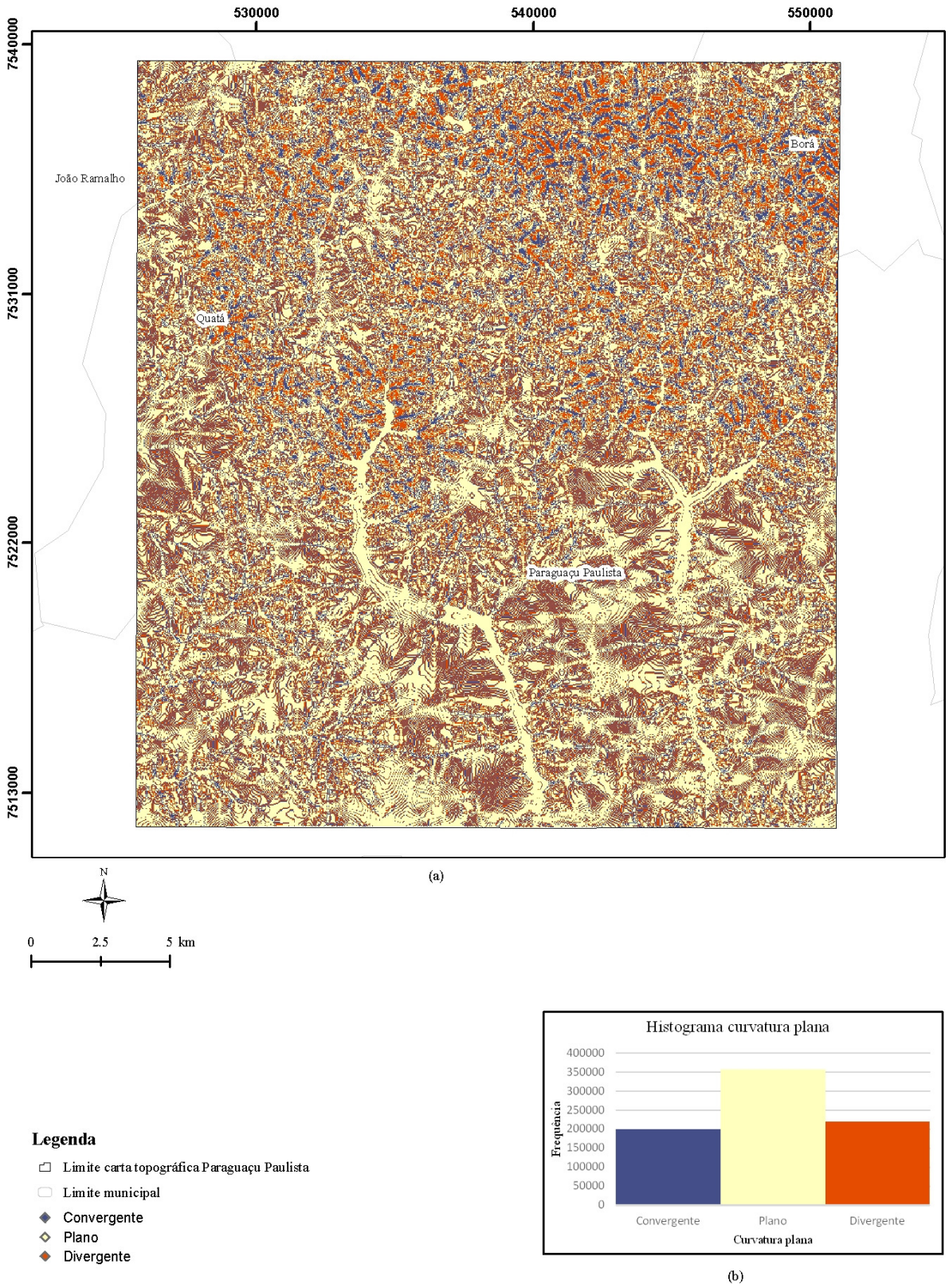


Figura 22. Mapa (a) e histograma (b) da variável curvatura plana recortada para a carta de Paraguaçu Paulista.

Analisando quantitativamente, as classes convergente e divergente possuem quantidade de ocorrência bem aproximadas (25,7 e 28,3% respectivamente), e com 46% a classe plano apresentou predominância comparado as outras classes de curvatura horizontal da carta.

Semelhante ao mapa da curvatura horizontal, a variável curvatura vertical (ou também denominada curvatura em perfil) também foi representada pela classe retilíneo (valores nulos de curvatura) com maior número de frequência, comparada as classes côncavo e convexo (Figura 23a). Em números, as classes côncavo e convexo representam um pouco mais que 50% de toda a área (Figura 23b), com a segunda classe apresentando maior representatividade próximos aos corpos d´água, e a classe retilíneo apresentando 46,6% de todos os valores de curvatura presentes na carta.

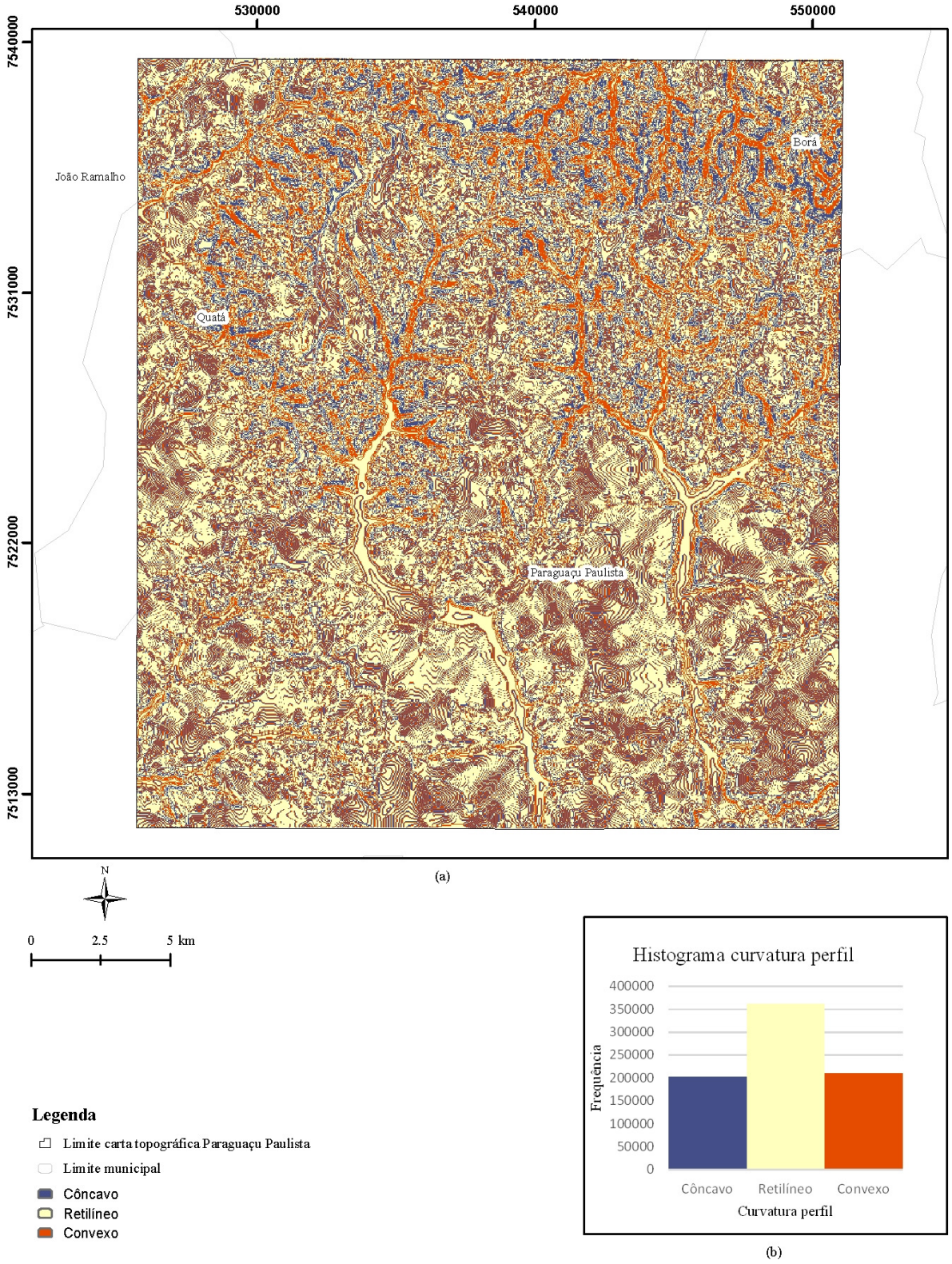


Figura 23. Mapa (a) e histograma (b) da curvatura perfil (vertical) da carta Paraguaçu Paulista.

Um dos fatores para a predominância das classes de curvatura plana e retilínea presentes nos mapas de curvatura horizontal e vertical de Paraguaçu Paulista pode ser explicado pela escala que foi realizado o estudo (1:50.000). A escala de semi-detalhe não demonstrou de maneira detalhada as

variações topográficas do terreno, isso possibilita dizer que mapas realizados em escalas maiores poderão apresentar variações mais acentuadas nos valores da curvatura, e consequentemente diferentes frequências de ocorrência de suas classes. O mesmo que foi explanado sobre as classes de curvatura da carta Quatá, onde trabalhos com escalas maiores provavelmente poderiam identificar maior variação nas classes de curvatura.

A porção norte da carta Paraguaçu Paulista, onde se limita com a carta Quatá está localizado o relevo de maior movimentação, com predominância da classe de declive montanhoso (Figura 24a). Ao contrário, no sul da carta de Paraguaçu, as classes suave ondulado e plano são mais perceptíveis.

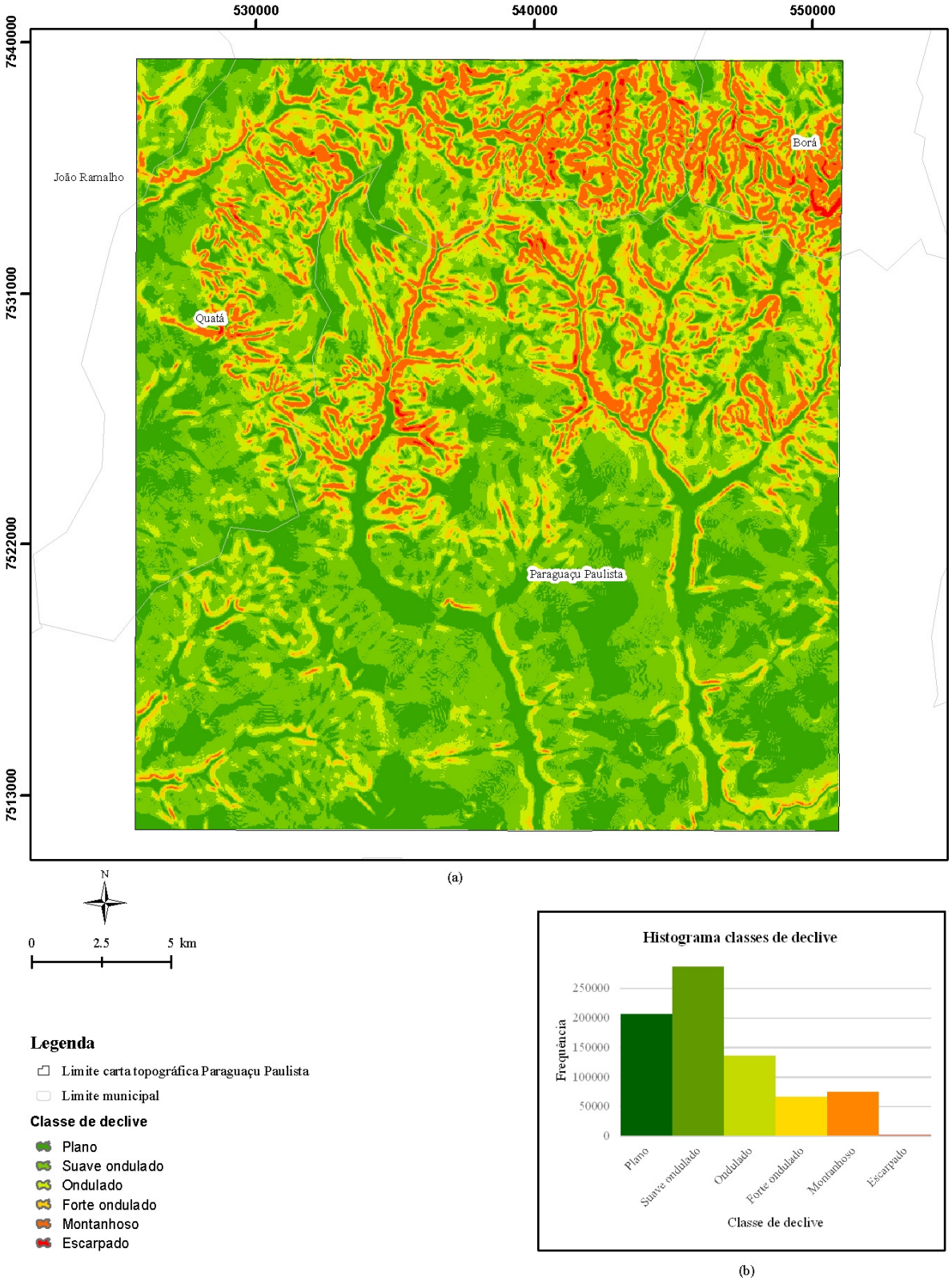


Figura 24. Mapa (a) e histograma (b) das classes de declive presentes na carta topográfica Paraguaçu Paulista.

De maneira geral, a carta de Paraguaçu paulista apresenta relevo suave, pela predominância da classe de declividade suave ondulado e plano (37 e 26,6% respectivamente). Movimentações do terreno mais abruptas, são representadas com menor expressão pelas classes escarpado, montanhoso e forte ondulado (0,2, 9,8 e 8,7% respectivamente).

Pela Figura 25a percebe-se onde estão os locais caracterizados pela menor e maior concentração de umidade. A região mais clara no norte, nordeste da carta Paraguaçu Paulista apresenta coloração mais clara, denotando menor quantidade de umidade, região essa onde o relevo é mais acidentado. No centro-sul estão presentes algumas áreas com concentração de umidade pouco superior, devido a presença de uma coloração mais escura.

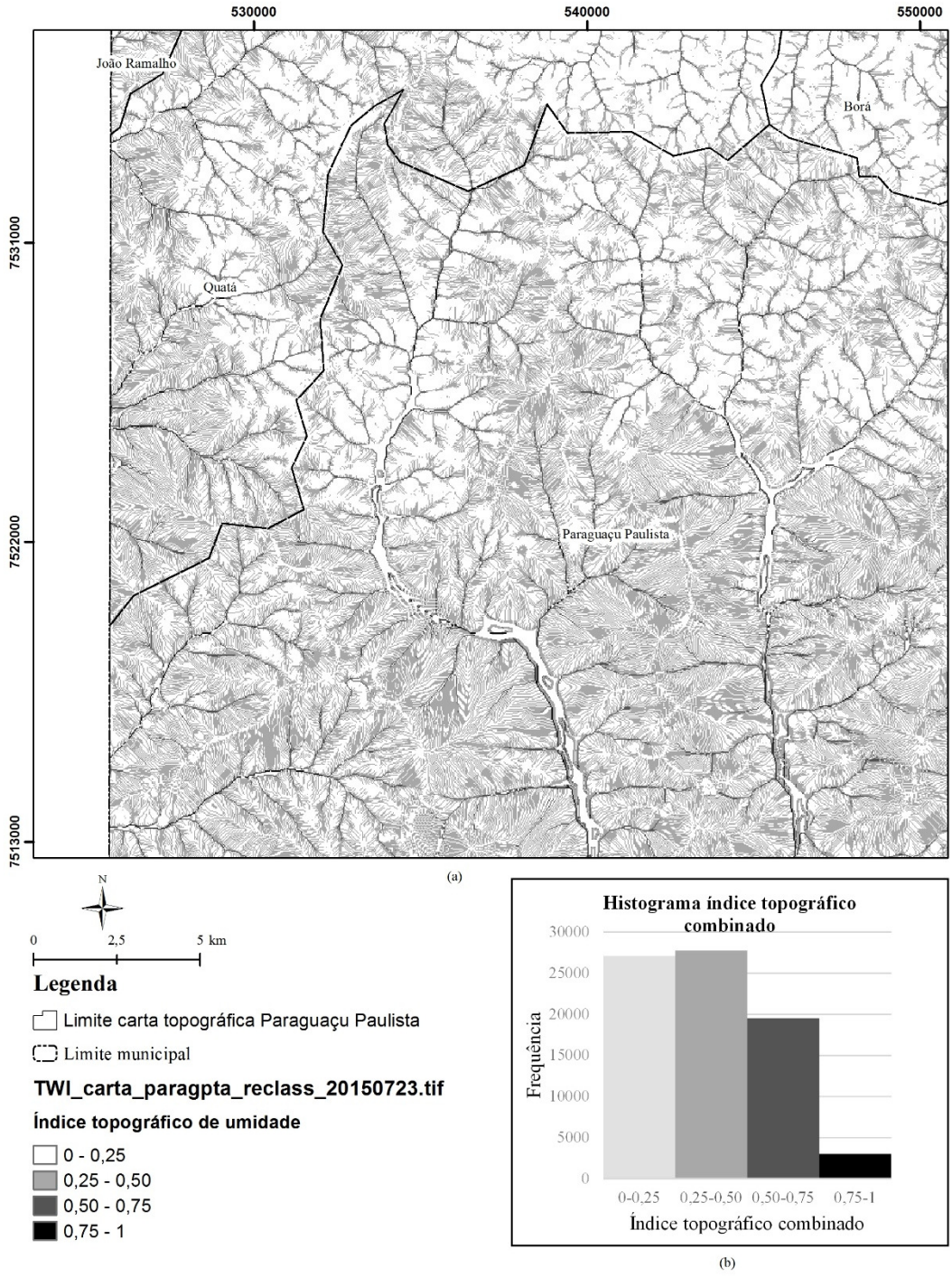


Figura 25. Mapa (a) e histograma (b) do índice topográfico combinado para a carta Paraguaçu Paulista.

Os locais com índice topográfico combinado elevados são muito escassos na carta de Paraguaçu Paulista, onde os índices de 0,5 a 1 representam apenas aproximadamente 4% da área total.

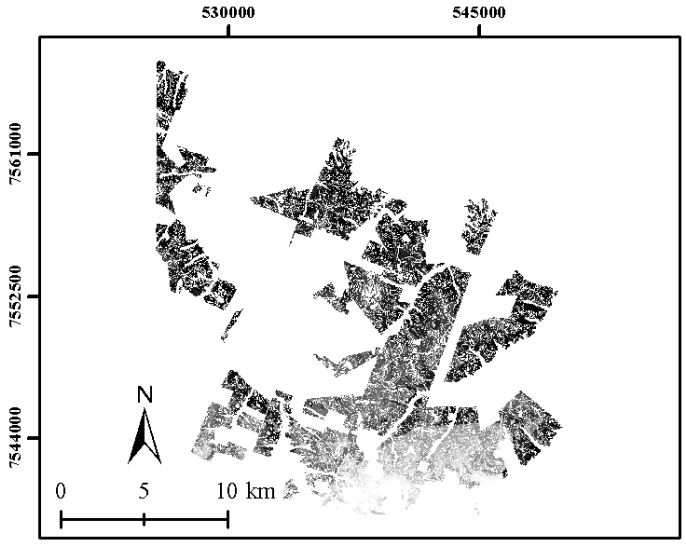
As classes do índice topográfico combinado de 0 – 0,5 abrangem quase que toda a carta, com aproximadamente 96%.

4.2 Cruzamento das variáveis morfométricas, altitude, geologia e pedologia

4.2.1 Elaboração da matriz de dados do recorte da área na carta Quatá para treinamento do modelo do mapa digital de solos

Os dados classificados apresentados na sessão anterior foram cruzados para elaboração de uma matriz única de dados com todas as informações.

A Figura 26a representa as imagens dos dados de altitude, geologia, morfometria e pedologia cruzados para a área recortada na carta Quatá, ou seja, cada pixel da imagem contém informação de cada uma dessas variáveis, além da informação espacial.



(a)

MDE	Geologia	Solo	OrientVert	CurvPlana	CurvPerfil	Decliv	TWI
340-445	Vale do Rio do Peixe	Argissolo Vermelho-Amarelo	Oeste	Plano	Retilineo	Plano	0,25-0,5
340-445	Vale do Rio do Peixe	Argissolo Vermelho-Amarelo	Norte	Plano	Convexo	Plano	0,25-0,5
340-445	Vale do Rio do Peixe	Argissolo Vermelho-Amarelo	Nordeste	Plano	Retilineo	Plano	0,25-0,5
340-445	Vale do Rio do Peixe	Argissolo Vermelho-Amarelo	Noroeste	Convergente	Convexo	Suave Ondulado	0,5-0,75
340-445	Vale do Rio do Peixe	Argissolo Vermelho-Amarelo	Nordeste	Convergente	Convexo	Suave Ondulado	0,5-0,75
340-445	Vale do Rio do Peixe	Argissolo Vermelho-Amarelo	Nordeste	Plano	Retilineo	Plano	0,5-0,75
340-445	Vale do Rio do Peixe	Argissolo Vermelho-Amarelo	Norte	Convergente	Convexo	Ondulado	0,5-0,75
340-445	Vale do Rio do Peixe	Argissolo Vermelho-Amarelo	Nordeste	Convergente	Convexo	Suave Ondulado	0,25-0,5

(b)

Figura 26. Representação espacial (a) e tabela (b) dos dados de altitude, geologia, morfometria e pedologia cruzados para o recorte na carta Quatá. OrientVert= Orientação das vertentes; CurvPlana=Curvatura Plana; CurvPerfil = Curvatura em perfil; Decliv = Declividade em percentagem; TWI = Índice topográfico combinado.

O produto matricial da combinação das variáveis apresentou 148.127 pixels, com 30x30 metros cada, ou seja, 13.331 hectares de área ocupada pela matriz de dados da carta Quatá. Cada pixel representa uma linha na Figura 24b, o que significa a manipulação de tabelas muito extensas.

Para importação desses dados no software de mineração de dados, foi necessário eliminar os caracteres especiais, acentuação e espaço entre as palavras (Tabela 03), bem como a padronização para leitura do software (Figura 27)

Tabela 03. Recorte da matriz de dados de altitude, geologia, morfometria e pedologia recortado para a carta Quatá utilizada para treinamento do modelo do mapa digital de solos.

MDE	Geologia	Solo	DirFluxo	CurvPlana	CurvPerfil	Decliv	TWI
340-445	Vale_do_Rio_do_Peixe	Argissolo_Vermelho-Amarelo	Oeste	Plano	Retilineo	Plano	0,25-0,5
340-445	Vale_do_Rio_do_Peixe	Argissolo_Vermelho-Amarelo	Norte	Plano	Convexo	Plano	0,25-0,5
340-445	Vale_do_Rio_do_Peixe	Argissolo_Vermelho-Amarelo	Nordeste	Plano	Retilineo	Plano	0,25-0,5
340-445	Vale_do_Rio_do_Peixe	Argissolo_Vermelho-Amarelo	Noroeste	Convergente	Convexo	Suave_Ondulado	0,5-0,75
340-445	Vale_do_Rio_do_Peixe	Argissolo_Vermelho-Amarelo	Nordeste	Convergente	Convexo	Suave_Ondulado	0,5-0,75
340-445	Vale_do_Rio_do_Peixe	Argissolo_Vermelho-Amarelo	Nordeste	Plano	Retilineo	Plano	0,5-0,75
340-445	Vale_do_Rio_do_Peixe	Argissolo_Vermelho-Amarelo	Norte	Convergente	Convexo	Ondulado	0,5-0,75
340-445	Vale_do_Rio_do_Peixe	Argissolo_Vermelho-Amarelo	Nordeste	Convergente	Convexo	Suave_Ondulado	0,25-0,5

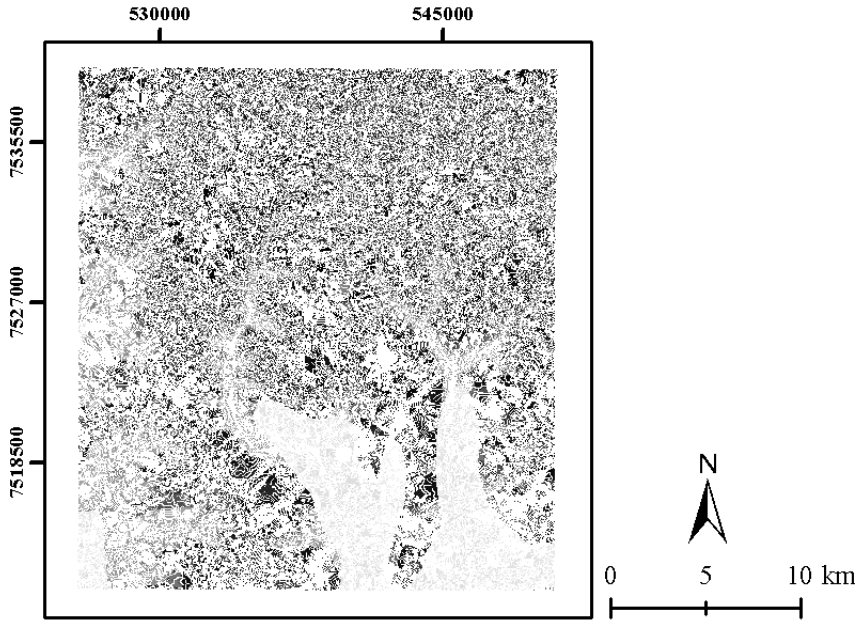
```
1  @relation Matriz_combine_quata_usina_20150730
2
3  @attribute TWI {0.25-0.5,0-0.25,0.75-1,0.5-0.75}
4  @attribute MDE {340-445,445-470,470-495,495-525,525-610}
5  @attribute CurvPerfil {Retilineo,Convexo,Concavo}
6  @attribute CurvPlana {Plano,Convergente,Divergente}
7  @attribute Declividad {Plano,Suave_Ondulado,Ondulado,Forte_Ondulado,Montanhoso,Escarpado}
8  @attribute OrientVertente {Oeste,Norte,Nordeste,Plano,Noroeste,Este,Sudoeste,Sudeste,Sul}
9  @attribute Geologia {Vale_do_Rio_do_Peixe,Marilia}
10 @attribute UNIMAP {ARGISSOLO_VERMELHO-AMARELO,LATOSSOLO_VERMELHO,ARGISSOLO_AMARELO,ARGISSOLO_VERMELHO,LATOS
11
12 @data
13 0.25-0.5,340-445,Retilineo,Plano,Plano,Oeste,Vale_do_Rio_do_Peixe,ARGISSOLO_VERMELHO-AMARELO
14 0-0.25,340-445,Convexo,Plano,Plano,Norte,Vale_do_Rio_do_Peixe,ARGISSOLO_VERMELHO-AMARELO
15 0.75-1,340-445,Retilineo,Plano,Plano,Nordeste,Vale_do_Rio_do_Peixe,ARGISSOLO_VERMELHO-AMARELO
16 0.25-0.5,340-445,Convexo,Plano,Plano,Plano,Vale_do_Rio_do_Peixe,ARGISSOLO_VERMELHO-AMARELO
17 0.25-0.5,340-445,Convexo,Plano,Plano,Norte,Vale_do_Rio_do_Peixe,ARGISSOLO_VERMELHO-AMARELO
18 0-0.25,340-445,Convexo,Convergente,Suave_Ondulado,Noroeste,Vale_do_Rio_do_Peixe,ARGISSOLO_VERMELHO-AMARELO
19 0-0.25,340-445,Convexo,Convergente,Suave_Ondulado,Nordeste,Vale_do_Rio_do_Peixe,ARGISSOLO_VERMELHO-AMARELO
20 0.75-1,340-445,Convexo,Plano,Plano,Nordeste,Vale_do_Rio_do_Peixe,ARGISSOLO_VERMELHO-AMARELO
21 0.25-0.5,340-445,Retilineo,Plano,Plano,Nordeste,Vale_do_Rio_do_Peixe,ARGISSOLO_VERMELHO-AMARELO
22 0.25-0.5,340-445,Convexo,Plano,Plano,Plano,Vale_do_Rio_do_Peixe,ARGISSOLO_VERMELHO-AMARELO
23 0-0.25,340-445,Convexo,Convergente,Plano,Noroeste,Vale_do_Rio_do_Peixe,ARGISSOLO_VERMELHO-AMARELO
24 0-0.25,340-445,Convexo,Plano,Suave_Ondulado,Norte,Vale_do_Rio_do_Peixe,ARGISSOLO_VERMELHO-AMARELO
25 0-0.25,340-445,Convexo,Convergente,Ondulado,Norte,Vale_do_Rio_do_Peixe,ARGISSOLO_VERMELHO-AMARELO
26 0-0.25,340-445,Convexo,Plano,Ondulado,Norte,Vale_do_Rio_do_Peixe,ARGISSOLO_VERMELHO-AMARELO
27 0-0.25,340-445,Convexo,Plano,Suave_Ondulado,Norte,Vale_do_Rio_do_Peixe,ARGISSOLO_VERMELHO-AMARELO
28 0-0.25,340-445,Convexo,Plano,Suave_Ondulado,Nordeste,Vale_do_Rio_do_Peixe,ARGISSOLO_VERMELHO-AMARELO
29 0.75-1,340-445,Convexo,Convergente,Suave_Ondulado,Nordeste,Vale_do_Rio_do_Peixe,ARGISSOLO_VERMELHO-AMARELO
30 0.75-1,340-445,Convexo,Plano,Plano,Nordeste,Vale_do_Rio_do_Peixe,ARGISSOLO_VERMELHO-AMARELO
```

Figura 27. Recorte da matriz de dados de altitude, geologia, morfometria e pedologia recortado para a carta Quatá utilizada para treinamento do modelo no formato para importação no software de mineração.

O formato do arquivo de entrada para mineração de dados é composto por um cabeçalho onde a primeira linha descreve uma identificação da matriz e da 3ª até a 10ª linha estão os atributos que serão utilizados para predizer o atributo classe (UNIMAP, linha número 10), seguidos de seus valores. Abaixo dos atributos (a partir da linha 13) estão os valores de cada atributo.

4.2.2 Elaboração da matriz de dados do recorte da carta Paraguaçu Paulista para teste do modelo do mapa digital de solos

Para a carta de Paraguaçu Paulista foram cruzadas as mesmas variáveis, exceto a de pedologia que será predita pelo modelo (Figura 28a).



(a)

CurvPerfil	CurvPlana	DecPer	DirFluxo	Geologia	MDE	TWI
Convexo	Plano	Plano	Sudoeste	Vale do Rio do Peixe	495-525	0,25-0,50
Convexo	Plano	Suave Ondulado	Este	Vale do Rio do Peixe	470-495	0,25-0,50
Convexo	Plano	Plano	Este	Vale do Rio do Peixe	470-495	0,25-0,50
Retilíneo	Plano	Plano	Oeste	Vale do Rio do Peixe	470-495	0,50-0,75
Convexo	Plano	Plano	Oeste	Vale do Rio do Peixe	470-495	0,25-0,50
Convexo	Convergente	Suave Ondulado	Oeste	Vale do Rio do Peixe	495-525	0,50-0,75
Retilíneo	Plano	Suave Ondulado	Oeste	Vale do Rio do Peixe	495-525	0,50-0,75
Côncavo	Plano	Suave Ondulado	Oeste	Vale do Rio do Peixe	495-525	0,50-0,75

(b)

Figura 28. Representação espacial (a) e tabela (b) dos dados de altitude, geologia e morfometria cruzados recortados para a carta de Paraguaçu Paulista.

Pelo fato da matriz que testou o modelo compreender uma carta topográfica de escala 1:50.000 inteira, consequentemente seu tamanho será maior que o recorte na carta Quatá utilizado para treinamento do modelo. No total o raster apresentou 571.251 pixels com 900 metros quadrados, representando uma área de 51.400 hectares.

O mesmo procedimento de limpeza e padronização dos dados também foi realizado para a tabela de 571.251 linhas utilizada para testar o modelo. A Tabela 04 demonstra as informações das variáveis cruzadas já organizadas para padronização e importação no software de mineração de dados.

Tabela 04. Recorte da matriz de dados das variáveis morfométricas da carta Paraguaçu Paulista

CurvPerfil	CurvPlana	DecPer	DirFluxo	Geologia	MDE	TWI
Convexo	Plano	Plano	Sudoeste	Vale_do_Rio_do_Peixe	495-525	0,25-0,50
Convexo	Plano	Suave_Ondulado	Este	Vale_do_Rio_do_Peixe	470-495	0,25-0,50
Convexo	Plano	Plano	Este	Vale_do_Rio_do_Peixe	470-495	0,25-0,50
Retilineo	Plano	Plano	Oeste	Vale_do_Rio_do_Peixe	470-495	0,50-0,75
Convexo	Plano	Plano	Oeste	Vale_do_Rio_do_Peixe	470-495	0,25-0,50
Convexo	Convergente	Suave_Ondulado	Oeste	Vale_do_Rio_do_Peixe	495-525	0,50-0,75
Retilineo	Plano	Suave_Ondulado	Oeste	Vale_do_Rio_do_Peixe	495-525	0,50-0,75
Concavo	Plano	Suave_Ondulado	Oeste	Vale_do_Rio_do_Peixe	495-525	0,50-0,75

Como a matriz de dados utilizada para testar o modelo não contém as classes de solo, o caractere que o Weka 3.7 utiliza para substituir valores faltantes da classe a ser predita é o ponto de interrogação (“?”). De acordo com a Figura 29 pode-se verificar a estrutura da matriz e seus valores para testar o modelo do mapa digital de solos.

```
@relation Book1

@attribute TWI {'0,25-0,50','0,50-0,75'}
@attribute MDE {495-525,470-495}
@attribute CurvPerfil {Convexo,Retilineo,Concavo}
@attribute CurvPlana {Plano,Convergente}
@attribute Declividad {Plano,Suave_Ondulado}
@attribute OrientVertente {Sudoeste,Este,Oeste}
@attribute Geologia {Vale_do_Rio_do_Peixe}
@attribute UNIMAP {ARGISSOLO_VERMELHO-AMARELO,LATOSSOLO_VERMELHO,ARGISSOLO_AMARELO,ARGISSOLO_VERMELHO

@data
'0,25-0,50',495-525,Convexo,Plano,Plano,Sudoeste,Vale_do_Rio_do_Peixe,?
'0,25-0,50',470-495,Convexo,Plano,Suave_Ondulado,Este,Vale_do_Rio_do_Peixe,?
'0,25-0,50',470-495,Convexo,Plano,Plano,Este,Vale_do_Rio_do_Peixe,?
'0,50-0,75',470-495,Retilineo,Plano,Plano,Oeste,Vale_do_Rio_do_Peixe,?
'0,25-0,50',470-495,Convexo,Plano,Plano,Oeste,Vale_do_Rio_do_Peixe,?
'0,50-0,75',495-525,Convexo,Convergente,Suave_Ondulado,Oeste,Vale_do_Rio_do_Peixe,?
'0,50-0,75',495-525,Retilineo,Plano,Suave_Ondulado,Oeste,Vale_do_Rio_do_Peixe,?
'0,50-0,75',495-525,Concavo,Plano,Suave_Ondulado,Oeste,Vale_do_Rio_do_Peixe,?
```

Figura 29. Recorte da matriz de dados de altitude, geologia e morfometria para a carta Paraguaçu Paulista utilizada para testar o modelo no formato para importação no software de mineração.

4.3 Análise dos dados e elaboração do modelo preditivo de solos

A matriz de dados da folha Quatá foi importada no software de mineração de dados Weka 3.7. Como na amostra de solos coletada na carta de Paraguaçu Paulista não consta as unidades de mapeamento NEOSSOLO LITÓLICO E GLEISSOLO HÁPLICO, esses foram retirados da matriz de dados.

A partir da matriz de dados, foram separados 70% para treinar o modelo e 30% para validá-lo, posteriormente foi realizado um pré-processamento, consistindo na discretização dos valores das variáveis e balanceamento das classes. E por último o modelo foi testado utilizando a matriz de dados sem classes de solo da carta Paraguaçu Paulista.

A separação dos dados de treinamento e validação foi realizada pelo próprio programa de mineração, onde os dados foram selecionados de maneira aleatória, compreendendo todos os atributos e valores sem repetição e sobreposição. Essa metodologia foi utilizada visando um maior aprendizado pelo programa (HASTIE et al., 2009).

A discretização consiste na transformação dos atributos em intervalos de valores, pois alguns algoritmos de classificação e agrupamentos (*clustering*) não são executados em atributos numéricos, apenas com atributos nominais (WITTEN et al., 2011). Nesse caso, no próprio ArcGIS os dados já foram reclassificados obedecendo o padrão identificado pelos algoritmos do Weka.

O balanceamento de classes torna-se necessário quando deseja-se predizer um número considerável de classes e a quantidade delas na base de dados são desproporcionais. A ferramenta aplica um filtro que controla a proporção de exemplos positivos/negativos no conjunto de dados de treinamento. É realizado um balanceamento do conjunto de dados por meio de uma amostragem com reposição, sempre mantendo o número de exemplos no conjunto de treinamento constante (WITTEN et al., 2011) O balanceamento é realizado dentro de um intervalo entre 0 e 1, onde valores próximos a zero a distribuição final será similar a inicial, e valores próximos a 1 a distribuição final será próximo da balanceada. No presente estudo o modelo foi elaborado com o conjunto de treinamento balanceado com o valor 0,5 e sem balanceamento, valor igual a zero (Figura 30).

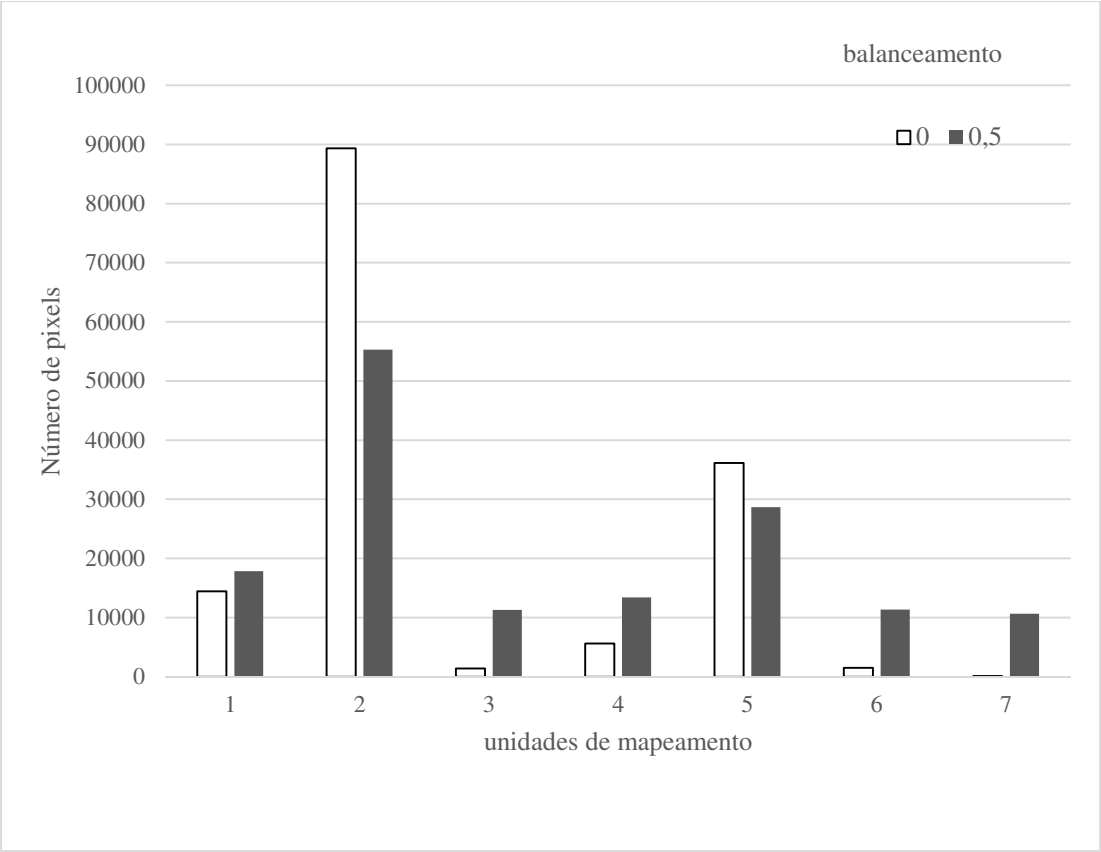


Figura 30. Distribuição dos pixels por unidades de mapeamento do recorte na folha Quatá com e sem balanceamento: 1– ARGISSOLO VERMELHO-AMARELO, 2 – LATOSSOLO VERMELHO, 3 – ARGISSOLO AMARELO, 4 – LATOSSOLO VERMLEHO-AMARELO, 5 – ARGISSOLO VERMELHO, 6 – LATOSSOLO AMARELO, 7 – NEOSSOLO QUARTZARÊNICO.

Analisando os dados sem aplicação do balanceamento, as classes LATOSSOLO VERMELHO e ARGISSOLO VERMELHO representam 84% de todos os dados, enquanto que as classes ARGISSOLO AMARELO, NEOSSOLO QUARTZARÊNICO e LATOSSOLO AMARELO representam no total 2% do número total de instâncias.

Após o balanceamento das classes houve uma diminuição do número de pixels (34 mil pixels para o LATOSSOLOO VERMELHO por exemplo) que classes sobreamostradas apresentavam (LATOSSOLO VERMELHO e ARGISSOLO VERMELHO). Parte desses dados pertencentes as classes com maior número de dados foram remanejados para as classes subamostradas. Com isso, classes como NEOSSOLO QUARTZARÊNICO, LATOSSOLO AMARELO e ARGISSOLO AMARELO tiveram maior representatividade no modelo de classificação (totalizando aproximadamente 22%).

Realizado pré-processamento, os modelos foram treinados, avaliados e testados. Todos os modelos avaliados foram construídos pelo algoritmo de classificação C4.5 (no Weka 3.7 é nomeado como J48) para geração de árvore de decisão.

No presente estudo o balanceamento das classes proporcionou a interação das classes com menor quantidade de dados no modelo gerado, porém, a acurácia geral do modelo diminuiu (Tabela 05). Enquanto que sem balanceamento de classe o modelo teve uma acurácia de aproximadamente 67%, com balanceamento de 0,5 seu valor decresceu para 58,27.

Para melhorar a capacidade dos modelos em predizer os solos BAILEY et al. (2003) sugerem que sejam desconsideradas as classes com área menor do que 5% da área total mapeada.

Tabela 05. Acurácia geral do modelo com balanceamento e sem balanceamento das classes

Balanceamento de classes	**0,0**	**0,5**
Acurácia geral do modelo (%)	67,67	58,27
Coeficiente Kappa	0,3145	0,2688

A acurácia do modelo é uma porcentagem do valor de instâncias corretamente classificadas pelo total de instâncias que o modelo gerou. Na maioria dos casos o classificador possui uma boa acurácia para a classe majoritária, mas uma acurácia baixa para a classe minoritária. A diminuição do valor da acurácia no modelo de dados com balanceamento igual a 0,5 ocorre pois classes que não tinham entrado no modelo passam a ter maior representatividade, consequentemente há um aumento das classificação incorretas. Juntamente com a relação de erros e acertos o coeficiente Kappa também diminuiu para o modelo gerado com balanceamento das classes.

Analisando a acurácia específica de cada classe, as unidades de mapeamento ARGISSOLO VERMELHO-AMARELO, ARGISSOLO AMARELO, LATOSSOLO VERMELHO-AMARELO, ARGISSOLO VERMELHO e LATOSSOLO VERMELHO apresentaram um acréscimo no valor de acurácia quando comparado os dados sem e com balanceamento, ao contrário da avaliação geral do modelo (acurácia geral) que teve um decréscimo. Já a acurácia da unidade de mapeamento com grande representatividade diminuiu (LATOSSOLO VERMELHO), devido à deflação que o balanceamento de classes condiciona.

Tabela 06. Acurácia[1] do classificador para cada classe de solo realizado com balanceamento e sem balanceamento.

Classe de solo	Balanceamento	
	0	0,5
ARGISSOLO VERMELHO-AMARELO	0,715	0,725
LATOSSOLO VERMELHO	0,721	0,710
ARGISSOLO AMARELO	0,797	0,833
LATOSSOLO VERMELHO-AMARELO	0,717	0,729
ARGISSOLO VERMELHO	0,736	0,739
LATOSSOLO AMARELO	0,842	0,880
NEOSSOLO QUARTZARENICO	0,831	0,827

[1] Acurácia medida pela área embaixo da curva (sigla em inglês AUC).

A curva ROC (*receiver operating characteristic*) é uma medida para avaliação de classificadores em mineração de dados. Utilizada na situação onde o modelo de aprendizagem tenta retirar amostras dos melhores acertos das instâncias testadas. A performance do classificador é determinada sem levar em conta a distribuição das classes (WITTEN et al., 2011). O gráfico da curva ROC (Figura 31) permite visualizar os valores fora da curva entre as taxas de positivos-verdadeiros e positivos falsos[2] de vários modelos de um classificador. Pela medição da área embaixo da curva do modelo, e quanto maior a área, maior será a qualidade do classificador.

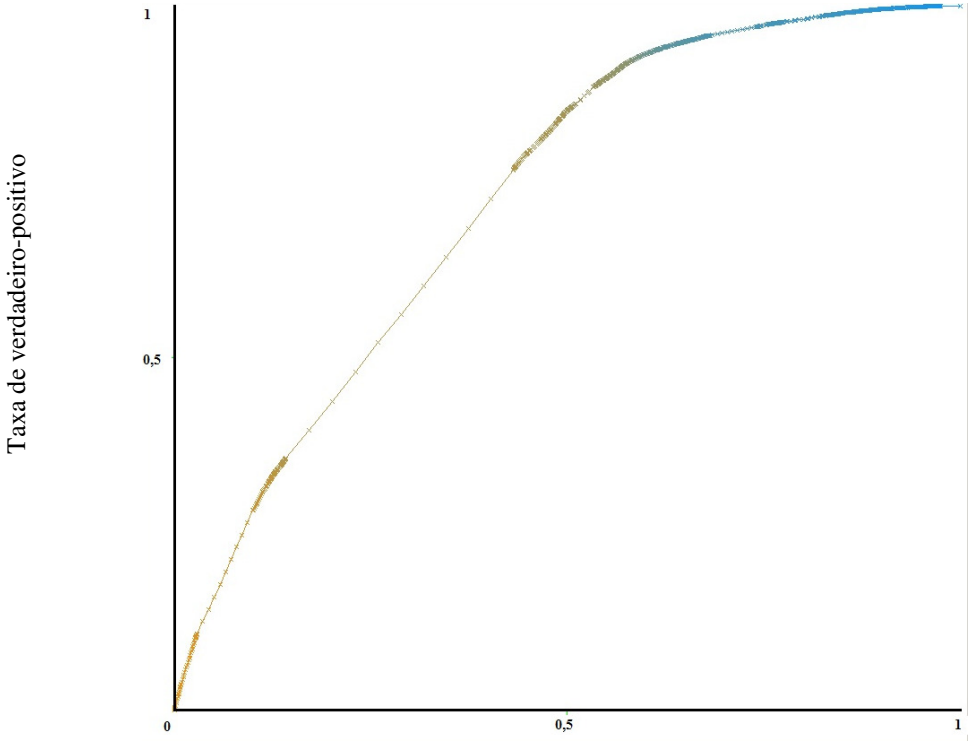


Figura 31. Exemplo de gráfico da curva ROC da unidade de mapeamento LATOSSOLO VERMELHO.

[2] A taxa de verdadeiros-positivos corresponde à medida de sensitividade (ou revocação). É definida como a porcentagem de elementos corretamente classificados como positivos pelo modelo dentre todos os positivos reais. Já a taxa de falso-positivo é definida como a porcentagem de elementos erroneamente classificados como positivos pelo modelo dentre todos os negativos reais (WITTEN et al., 2011).

As variáveis que mais contribuíram para o modelo de mapa digital de solos estão representadas na Tabela 07, onde a litologia e a altitude aparecem como principais atributos na predição de classes de solo. A declividade e orientação da vertente alternaram a 3ª e 4ª posição no rank quando o modelo foi gerado com balanceamento e sem balanceamento das classes.

Tabela 07. Ordenamento dos atributos quanto à contribuição no modelo de mapa digital de solos, obtido pelo teste de entropia.

Rank	Atributo	
	Sem balanceamento	**Com balanceamento**
1	Geologia	Geologia
2	MDE	MDE
3	Declividade	OrientVertente
4	OrientVertente	Declividade
5	TWI	TWI
6	CurvPerfil	CurvPerfil
7	CurvPlana	CurvPlana

A variável geologia foi determinante na atribuição das unidades de mapeamento, como por exemplo toda litologia da unidade Serra Geral foi classificada como LATOSSOLO VERMELHO, não necessitando nenhuma outra informação para o algoritmo chegar a classe atribuída.

A altitude foi a segunda variável que auxiliou o modelo na predição das unidades de mapeamento, seja no modelo gerado com classes com balanceamento, seja sem balanceamento. Enquanto que as variáveis de curvatura foram pouco determinantes para o modelo classificar a unidade de mapeamento.

4.4 Mapa digital das classes de solo e validação por pontos obtidos em campo

A Figura 32 representa o mapa digital de classes de solo obtido pelo modelo gerado sem balanceamento das classes. Nesse caso, as classes que apresentavam uma porcentagem muito pequena de valores comparada ao total utilizada para o modelo, foram suprimidas, pois não foram preditas por nenhum atributo.

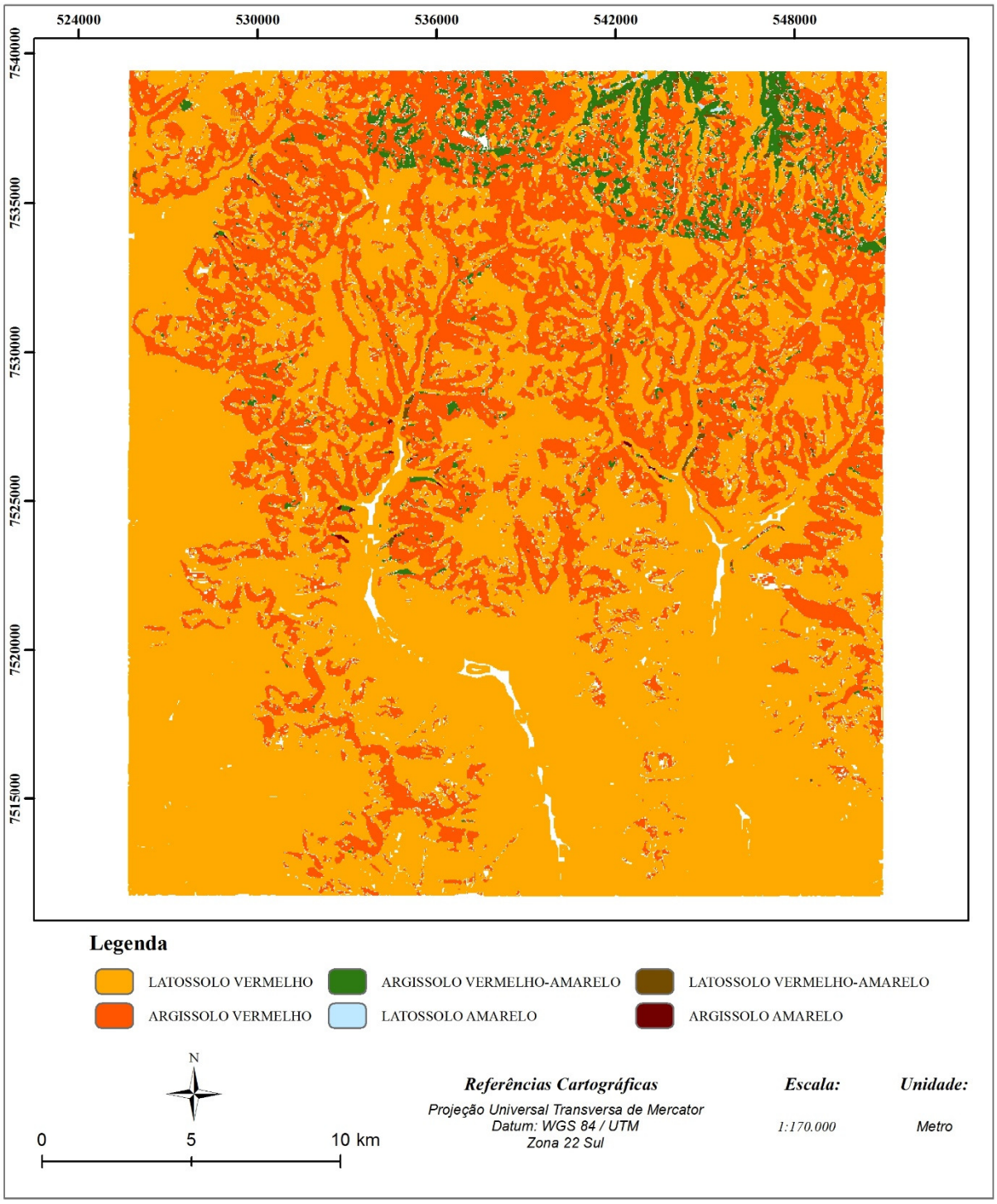


Figura 32. Mapa digital das classes de solo da carta de Paraguaçu Paulista, obtido pelo modelo sem balanceamento das classes.

A classe de solo LATOSSOLO VERMELHO representou aproximadamente 65% do total de classes preditas. A classe apresentou predominância nas áreas de declive plano e suave ondulado, aproximadamente 78% da unidade de mapemanto LATOSSOLO VERMELHO está inserida em locais onde o terreno é caracterizado por declives poucos acentuados (declividade plano e suave

ondulado). Na porção norte da carta, onde está presente a formação geológica Marília e também declividades mais acentuadas, a classe LATOSSOLO VERMELHO teve pouca representatividade. Nos locais onde a classe de declive foi classificada como Montanhoso e Escarpado a classe de solo com maior abrangência na carta representou apenas 4,5%, e dentro da formação geológica Marília a classe LATOSSOLO VERMELHO apresentou 4,6% de representatividade.

A segunda unidade de mapeamento com maior número de valores no mapa foi a ARGISSOLO VERMELHO, com aproximadamente 31% de toda a amostra do modelo. Sua localização na carta foi predominantemente em áreas de declive montanhoso, porém está distribuída por toda a área do mapa. A segunda classe de solo com maior representatividade comparada a todas as outras unidades de mapeamento, apresenta uma amostra de 78% localizada nas classes de declive Ondulado, Forte Ondulado, Montanhoso e Escarpado.

Com porcentagem igual a 3% de todas as classes, a unidade ARGISSOLO VERMELHO-AMARELO foi predita predominantemente (75%) onde a formação geológica é representada pela unidade de Marília, e também apresentou representatividade (54%) nas áreas com classes de declive forte ondulado e montanhoso.

O restante das classes somaram 0,65% de todas as preditas. As classes ARGISSOLO AMARELO, LATOSSOLO AMARELO E LATOSSOLO VERMLEHO-AMARELO estão distribuídas principalmente junto com a formação geológica de Marília, em altitudes médias a elevadas e declive ondulado a montanhoso.

Segundo PRADO (2007) a paisagem representa a ação combinada de diversos fatores de formação do solo, um desses é o relevo, que justifica a distribuição dos solos obtidos pelo modelo na região de Quatá. Nas superfícies mais antigas e estáveis da paisagem (relevo plano ou suavemente ondulado) geralmente predominam os Latossolos, e em superfícies semelhantes, porém mais recentes há presença dos Argissolos. Em menor expressão, próximo aos corpos d'água os Neossolos predominam.

A avaliação da acurácia geral do mapa após sua validação com pontos coletados em campo, foi de 49,25% e o índice Kappa de 0,16, o que demonstra baixa relação da concordância observada pela esperada.

Esse baixo valor de concordância entre o modelo gerado com a realidade é explicado pela distribuição das amostras de solos utilizadas para validação, representar áreas benéficas ao plantio de cana-de-açúcar. Os pontos com informações pedológicas foram determinados segundo estudo de solos para uma usina de cana-de-açúcar, o que pode representar apenas locais de interesse para essa pesquisa da usina, e não abragendo todos os locais da paisagem mais representativos para geração e validação de mapas digitais de solo.

Semelhante ao mapa digital elaborado sem balanceamento das classes, o modelo elaborado a partir das classes com balanceamento igual a 0,5 apresentou predominância da classe LATOSSOLO VERMELHO e as classes LATOSSOLO AMARELO, ARGISSOLO AMARELO e NEOSSOLO QUARTZARÊNICO obtiveram menor representatividade, porém houve um aumento no número de pixels das classes pouco predominantes e uma diminuição no número da amostra das classes com maior representatividade (Figura 33).

O LATOSSOLO VERMELHO representa cerca de 54% de todas as classes, com predominância nas formações geológicas Serra Geral e Vale do Rio do Peixe, e em relevo suave ondulado. A próxima classe com representatividade, é a ARGISSOLO VERMELHO, abrangendo cerca de 33% de todo o mapa. Essa classe está presente por todo o mapa, com maior concentração na região central e norte, principalmente em declives mais acentuados (de forte ondulado a escarpado).

Representando aproximadamente 6% das unidades mapeadas, a classe ARGISSOLO VERMELHO-AMARELO está presente principalmente dentro da formação geológica Marília com 59% e nos locais onde o terreno é mais inclinado (classes de declive ondulado e montanhoso), com 61 %.

As classes de solo restantes, com menor representatividade no mapa digital somaram aproximadamente 6,5% do total de unidades de mapeamento presente no mapa. As classes estão presentes em praticamente toda a imagem, com predominância no norte (divisa com a carta topográfica de Quatá), onde as variações altimétricas e de declive são acentuadas e em altitudes moderadas pelo mapa.

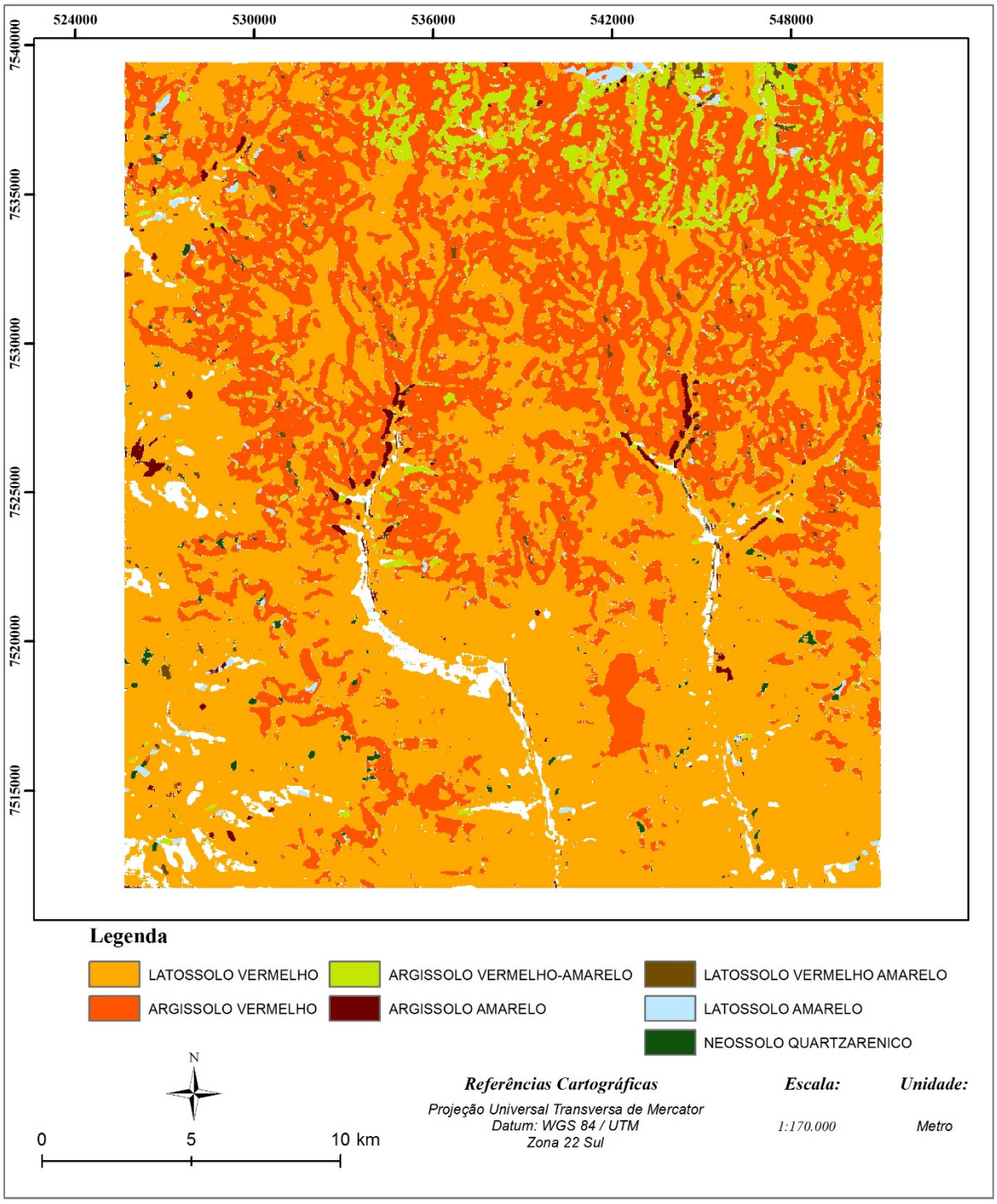


Figura 33. Mapa digital das classes de solo da carta de Paraguaçu Paulista, obtido pelo modelo com balanceamento de 0,5 das classes.

A acurácia geral medida após validação com os pontos de solos coletados em campo, para o mapa digital gerado com balanceamento das classes foi de aproximadamente 45% e índice Kappa igual a 0,13. O índice representa um baixo grau de concordância entre as classes existentes no terreno

com as que foram preditas pelo modelo. LANDIS e KOCH (1977) caracterizaram diferentes escalas de valores para o índice kappa e seus respectivos graus de concordâncias. Valores maior que 0,75 representam um excelente grau de concordância, valores abaixo de 0,40 são tidos como grau de concordância fraco e valores entre 0,40 e 0,75 representam média a boa concordância.

Analisando o comportamento dos modelos, de maneira geral, tem-se que a grande representatividade das classes LATOSSOLO VERMELHO e ARGISSOLO VERMELHO presente na base de dados de entrada dos modelos influenciou na predição das classes de solo. Pelo Gráfico 02 percebe-se a predominância das classes que tiveram maior representatividade nos mapas digitais produzidos.

Com relação as variáveis morfométricas e atributos utilizados para geração do modelo, o atributo geologia foi o mais importante na distribuição das unidades de mapeamento no mapa, seguido das classes de altitude e declividade.

Os valores muito baixos obtidos no índice de concordância Kappa, foram devido a elaboração do modelo com dados da carta Quatá e aplicação em outra área diferente da avaliada pelo modelo, totalmente nova. Porém apesar desses valores baixos de concordância, o modelo compreendeu a relação dos atributos/variáveis morfométricas com a classe pedológica, ou seja, o modelo treinado na carta Quatá e testado em área diferente (carta Paraguaçu Paulista) apresentou concordância de suas classes de solo preditas com os atributos do relevo e geologia.

A utilização de extensa rede de dados de campo fornece subsídio para testar outras inúmeras hipóteses, bem como avaliar novos métodos, algoritmos e modelos. Portanto, a pedometria ainda é um campo que está crescendo e evoluindo, porém, ainda necessitando dos métodos tradicionais e qualitativos de classificação de solos. O mapeamento digital de solo satisfatório necessita das pesquisas de um pedólogo, seja para entrada de dados no modelo, identificando atributos necessários para serem inseridos no modelo, como também determinação da escala de trabalho, quantificação e qualificação das amostras necessárias, e percepção das unidades de mapeamento na paisagem.

5 CONCLUSÕES

- Com a utilização de técnicas de mineração de dados aliada às ferramentas existentes nos Sistemas de Informações Geográficas foi possível elaborar um mapa digital de classes de solo em nível de reconhecimento, utilizando variáveis geomorfométricas, geologia e mapas pedológicos pré existentes.
- Como constatados em outros estudos, a geologia, a altitude e a declividades constituíram em variáveis fundamentais na metodologia de mapeamento digital de solo.
- Em áreas com predominância de relevo plano a suave ondulado, a diferenciação das variáveis morfométricas é menor, principalmente se a informação altimétrica for oriunda de levantamento topográficos em escalas grandes, a partir de 1:50000.
- A manipulação de grandes matrizes de dados dificulta a modelagem de mapas digitais de solos, necessitando amplo poder computacional e prejudicando nas tentativas de novas métodos para aprimoramento do modelo.
- Como as informações pedológicas disponíveis para a área teste (Folha Quatá) são provenientes de um levantamento pedológico específico e orientado para áreas com solo e relevo mais adequados para plantio de cana de açúcar, provavelmente a área teste não representou toda a variabilidade de combinações "solo x relevo", exigidas nas análises de predição de classes de solo pelos algoritmos empregados. Isso pode ter explicado os baixos valores do índica Kappa obtidos.
- O mapeamento digital de solos constitui-se em uma ferramenta essencial para a geração de mapas pedológicos, principalmente no nível de reconhecimento, possibilitando um ganho significativo em termos de tempo e densidade amostral, em grandes áreas. Novos recursos metodológicos, como o emprego de autômatos celulares, poderão contribuir significativamente para o avanço do mapeamento digital de solos.
- O avanço das pesquisas para aprimoramento e criação de novos métodos para elaboração dos mapeamentos digitais de solo é necessário, porém não menos importante e ainda indispensável, é a presença de um pedólogo experiente para auxílio da pesquisa seja para entrada de dados no modelo, identificação dos atributos necessários para serem inseridos no modelo, como também determinação da escala de trabalho, quantificação e qualificação das amostras necessárias, e percepção das unidades de mapeamento na paisagem.

6 REFERÊNCIAS BIBLIOGRÁFICAS

ALMEIDA, F. F. M. Fundamentos geológicos do relevo paulista. **Boletim** de Geologia do Estado de São Paulo. – IGC, n. 41, p. 169-263, 1964.

AMO, SANDRA DE. Técnicas de mineração de dados. **In**: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO. JORNADA DE ATUALIZAÇÃO EM INFORMÁTICA, 24. 2004, Salvador.

ARONOFF, S. **Geographic information systems: a management perspective.** Ottawa: WDL Publications, 295p., 1989.

BATISTA, G. E. A. P. A. Pré-processamento de dados em aprendizado de máquinas supervisionado. 2003. 204 p. **Tese** (Doutorado) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos.

BASGALUPP, M. P. LEGAL Tree: Um algoritmo genético multi-objetivo lexicográfico para indução de árvores de decisão. 2010. 94p. **Tese** (Doutorado). Instituto de Ciências Matemáticas e de Computação – ICM - Universidade de São Paulo. São Carlos

BEVEN, K. J.; KIRBY, M. J. A Physically based, variable contributing area model of basin hydrology. **Hydrological Sciences Bulletin**, 24, 43-69, 1979.

BONGIOVANNI, S. Uma abordagem de geologia de engenharia ao cenozoico da região de Paraguaçu Paulista. **Dissertação** (Mestrado em Geociências e Meio Ambiente) – Instituto de Geociências e Ciências Exatas, Universidade Estadual Paulista, Rio Claro, 1990. 102 p.

BONGIOVANNI, S. Caracterização geológica do município de Assis: a importância do estudo das coberturas cenozoicas. **Tese** (Doutorado em Geologia Regional) Instituto de Geociências e Ciências Exatas, Universidade Estadual Paulista, Rio Claro, 2008. 218 p.).

BRAZDIL, P. Construção de modelos de decisão a partir de dados, 1999. Disponível em: <http://www.nacc.up.pt/~pbrazdil/Ensino/ML/ModDecis.html>. Acesso em 01 junho 2013.

BREIMAN, L.; FRIEDMAN, J.H.; OLSHEN, R.A.; STONE, C. **Classification and regression trees**. Chapman and Hall. Wadsworth. 368 p. 1984.

BROWN, D.J. A historical perspective on soil-landscape modeling. **In**: Grunwald (Ed.), Environmental Soil-landscape modeling: Geographic information technologies and pedometrics. CRC Press, Boca Raton, Fl, p. 61-103, 2005.

BUI, E. A review of digital soil mapping in Australia. **In**: Lagacherie, P.; McBratney, A.B.; Voltz, M. (Eds). Digital soil mapping: an introductory perspective. Amsterdam. Elsevier. p. 25-39. 2007.

CATEN, A. ten. **Aplicação de componentes principais e regressões logísticas múltiplas em sistema de informações geográficas para a predição e o mapeamento digital de solos**. 2008. 128 p.. Dissertação (Mestrado em Ciência do Solo) – Universidade Federal de Santa Maria, Santa Maria, 2008.

CATEN, A.; DALMOLIN, R. S. D.; MENDONÇA-SANTOS, M. L.; GIASSON, E. **Mapeamento digital de classes de solos: características da abordagem brasileira**. Ciência Rural, vol. 42, n. 11. 2012. Pp. 1989-1997.

CEPAGRI – Centro de Pesquisas Meteorológicas e Climáticas Aplicadas a Agricultura. Disponível em: <http://www.cpa.unicamp.br/outras-informacoes/clima_muni_402.html>. Acesso em janeiro 2015.

CHAGAS, C. S. Mapeamento digital de solos por correlação ambiental e redes neurais em bacia hidrográfica do domínio de mar de morros. **Tese** (Doutorado em solos e nutrição de plantas). Universidade Federal de Viçosa, Viçosa. 2006.

COELHO, R.M.; LEPSCH, I.F.; MENK, J.R.F. Relações solo-relevo em uma encosta com transição arenito-basalto em Jaú (SP). **Revista Brasileira de Ciência do Solo**, v. 18. n.1, p. 125-137, 1994.

COHEN, J. Weighted kappa: nominal scale agrément with provision for scaled disagreement or partial credit. **Psychological Bulletin**. N. 70. PP. 213-220. 1968.

CONGALTON, R. A review of assessing the accuracy of classifications of remotely data. **Remote Sensing Environ**. 37, pp. 35-46, 1991.

CPRM. SERVIÇO GEOLÓGICO DO BRASIL. **Mapa Geológico do estado de São Paulo**. Minstério de Minas e Energia – Secretaria de Geologia, Mineração e Transformação Mineral. Brasília, 2006. Escala 1:750.000.

CPTI – Cooperativa de Serviços, Pesquisas Tecnológicas e Industriais. Diagnóstico da situação dos Recursos Hídricos da UGRH 17 – Médio Paranapanema: **Relatório Zero**. São Paulo, 1999.

CRIVELENTI, R.C. Mineração de dados para inferência da relação solo-paisagem em mapeamentos digitais de solos. **Dissertação** (Mestrado em Agricultura Tropical e Subtropical). Campinas: Instituto Agronômico, 2009, 107p.

EDMONDS, W.J. Soil Mapping and Survey. **In**: Chesworth (Ed.), Encycolpedia of Soil Science. Springer, Dordrecht, The Netherland, p. 670 – 673, 2008.N, R

EMBRAPA. Sistema Brasileiro de Classificação de Solos. Rio de Janeiro: Embrapa, 2006, 2ª ed. 412p.

ESRI – Environmental System Research Institute. ArcGIS. Version 9.3. Redlands: ESRI,2008.

FERNANDES, L.A. Mapa litoestratigráfico da parte oriental da bacia Bauru (PR, SP, MG), escala 1:100.000**. Boletim** Paranaense de Geociências. 2004. 55 p. pp. 53-66.

GESSLER, P.E.; MOORE, I.D.; MCKENZIE, N.J. RYAN, P.J. Soil-landscape modeling and spatial prediction of soil attributes. **International Journal of Geographical Information Systems**, vol. 9, n.4, p. 421-432, 1995.

HAN J.; KAMBERM. **Data Mining**: concepts and techniques. Morgan Kaufmann Publishers Inc. San Francisco, CA. 2nd edition. 2006.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J, H. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. New York: Springer. 744 p. 2a ed. 2009.

HENGL, T. Pedometric mapping: briding the gaps between conventional and pedometric approaches. **Tese de Doutorado**. University of Wageningen, Enschede. 214p. 2003.

HOLSHEIMER, M. & SIEBES, A. Data Mining: the search for knowledge in databases. **Relatório** CS-R9406. Amsterdam, Holanda: CWI, 1994.

HUDSON, B.D. The soil survey as a paradigma-based Science. **Soil Science Society of America Journal**, v.56, p.836-841, 1992.

IBGE. **Manual Técnico de Pedologia**, 2ª edição. Rio de Janeiro: Ministério do Planejamento, Orçamento e Gestão. Instituto Brasileiro de Geografia e Estatística – IBGE. Diretoria de Geociências. Coordenação de Recursos Naturais e Estudos Ambientais. Manuais Técnicos em Geociências, número 4, 2007, p.316.

IBGE – Instituto Brasileiro de Geografia e Estatística. **Cartas Topográficas do Mapeamento Sistemático Brasileiro**, escala 1:50.000. IBGE/DSG. Disponível em <ftp://geoftp.ibge.gov.br>. Acesso em: julho 2013.

IPPOLITI, R.G.A.; COSTA, L.M.; SCHAEFER, C.E.G.R.; FILHO, E.I.F.; GAGGERO, M.R.; SOUZA, E. Análise digital de terreno: Ferramenta na identificação de pedoformas em microbacia na região de "mar de morros " (MG). **Revista Brasileira de Ciência do Solo**. V. 29, n.2, p. 269-276, 2005.

IPT – Instituto de Pesquisas Tecnológicas do Estado de São Paulo. **Mapa geomorfológico do Estado de São Paulo**, escala 1:1.000.000. 2v. 1981.

ITC - International Institute for Aeropspace Survey and Earth Sciences. ILWIS 3.3 for Windows: User´s Guide. Enchede: ITC, 2001.

JAIN, A., D. –W. e MAO, J. Statistical pattern recognition: A review. In **IEEE Transactions on Pattern Analysis and Machine inteligence** (22): 4-37, 2000.

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

JENSEN, J.R. **Introductory digital image processing**: a remote sensing perspective. 2nd ed. Upper Saddle River: Prentice Hall, 1996. 318 p.

KRONKA, F. J. N.; NALON, M. A.; MATSUKUMA, C. K.; KANASHIRO, M.M.; YWANE, M. S. S.; PAVÃO, M.; DURIGAN, G.; LIMA, L. M. P. R.; GUILLAUMON, J. R.; BAITELLO, J. B.; BORGO, S. C.; MANETTI, L. A.; BARRADAS, A. M. F.; FUKUDA, J. C.; SHIDA, C. N.; MONTEIRO, C. H. B.; PONTINHA, A. A. S.; ANDRADE, G. G.; BARBOSA, O.; SOARES, A. P.; **Inventário florestal da vegetação natural do estado de São Paulo**. São Paulo: Secretaria do Meio Ambiente; Instituto Florestal; Imprensa Oficial, 2005. 200p.

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. (Eds.). Digital soil mapping: An introductory perspective. Amsterdam. Elsevier. p. 3-22. 2007.

LANDIS, J. R. e KOCH, G. G. The measurement of observer agreement for categorical data. **Biometrics**, v. 33, n. 1, p. 159-174, 1977.

LEPSCH, I. F. &BUOL, S.W. Investigations in an Oxisol-Ultisol toposequence in Sao Paulo State, Brazil. **Soil Science Society of America Journal**, vol.38, p. 491-496, 1974.

IBGE - Malha municipal digital do Brasil: situação em 2000 e 2010. Rio de Janeiro: IBGE. Disponível em <ftp://geoftp.ibge.gov.br/malhas_digitais/>. Acesso em julho 2013.

MILNE, G. Normal erosion as a fator in soil profile development. **Nature** 138: 541-548. 1936.

MMA – Ministério do Meio Ambiente. Biomas. Disponível em: <http://www.mma.gov.br>. Acesso em janeiro 2015.

ONODA, M. & EBECKEN, N. F. F. Implementação em Java de um algoritmo de árvore de decisão acoplado a um SGBD relacional. **In** Marta Mattoso & Geraldo Xexéo. Ed. SBBD. COPPE/UFRJ. 2001. pp. 55-64.

PRADO, H. **Pedologia fácil**: aplicações na agricultura. 2 ed. São Paulo: Edição do Autor, 2007 ISBN-13:9788590133025.

MENDONÇA-SANTOS, M. DE; SANTOS, DOS H. G. **Mapeamento digital de classes e atributos de solos – métodos, paradigmas e novas técnicas**. 2003. EMBRAPA. 17p. Rio de Janeiro. Documentos 55.

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

RANZANI, G. **Manual de levantamento de solos**. Edgard Blucher Ltda. (Ed.).

ROKACH, L. & MAIMON, O. **Data Mining with decision trees. Theory and applications**. World Scientific Publishing. 2014.

SARMENTO, E.C.. Comparação entre quatro algoritmos de aprendizagem de máquina no mapeamento digital de solos no Vale dos Vinhedos, RS, Brasil. Dissertação (Mestrado), Universidade Federal do Rio Grande do Sul, Porto Alegre, 2010.

SILVA, E. F. Comparação de mapas de solos produzidos em escalas e épocas distintas. **Tese** (Doutorado). Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2000.

SOUZA, ZM.; MARQUES JÚNIOR, J.; PEREIRA, G.T.; MOREIRA, L.F. Influência da pedoforma na variabilidade espacial de alguns atributos físicos e hídricos de um Latossolo sob cultivo de cana-de-açúcar. **Irriga**, v.9, n.1, p.1-11, 2004.

STORY, M.; CONGALTON, R.G. Accuracy assesment: A user´s perspective. **Photogrammetric Engineering and Remote Sensing**, v.61, p. 397-399, 1986.

WEKA. Waikato Environment for Knowledge Analysis. Disponível em: <http://www.cs.waikato.ac.nz/ml/weka>. Acesso em novembro 2014.

WITTEN IAN H.; EIBE, F.; HALL, M. A. Data mining: practical machine learning tools and techniques. – 3rd ed. Morgan Kaufmann, San Francisco. 2011, 629p.