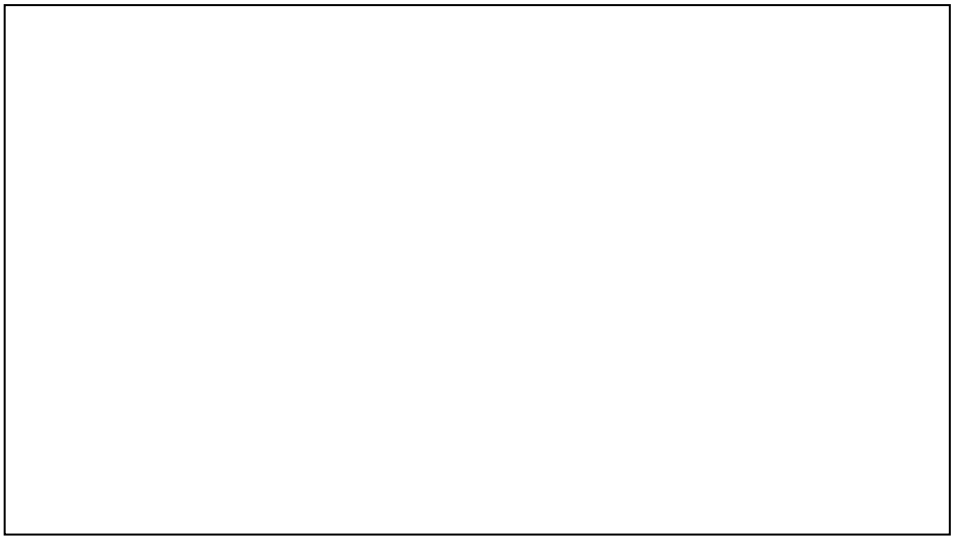
MINISTÉRIO DA CIÊNCIA E TECNOLOGIA
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

MINERAÇÃO DE PADRÕES DE MUDANÇA EM IMAGENS DE SENSORIAMENTO REMOTO

Marcelino Pereira dos Santos Silva

Tese de Doutorado em Computação Aplicada, orientada pelo Dr. Gilberto Câmara, aprovada em 3 de fevereiro de 2006.

INPE
São José dos Campos
2006



FOLHA DE APROVAÇÃO

"Ó profundidade das riquezas, tanto da sabedoria, como da ciência de Deus! Quão insondáveis são os seus juízos, e quão inescrutáveis os seus caminhos!
Porque, quem compreendeu a mente do Senhor? Ou quem foi seu conselheiro?
Ou quem deu primeiro a Ele, para que lhe seja recompensado?
Porque Dele e por Ele, e para Ele, são todas as coisas; glória, pois, a Ele eternamente. Amém."

Romanos 11:33-36

A meus pais,
MARIA AUGUSTA SILVA e
VIRGÍLIO PEREIRA DOS SANTOS (in memorian).

A minha família,
LENI ANDRADE BARROS DOS SANTOS,
DANIEL ANDRADE BARROS DOS SANTOS e
DAVI ANDRÉ ANDRADE BARROS DOS SANTOS.

AGRADECIMENTOS

Agradeço a Deus, o Altíssimo, pela oportunidade concedida e pela vitória alcançada.

Sou grato à minha família, que sempre me apoiou e incentivou em todos os momentos.

Minha gratidão ao Dr. Gilberto Câmara, pela competente orientação e preciosos conhecimentos transmitidos.

Agradeço também à Dra. Isabel Escada pela imprescindível colaboração neste trabalho, e ao Dr. Miguel Monteiro, Dr. Dalton Valeriano e Ricardo Cartaxo pelas relevantes discussões e sugestões.

Sou grato aos membros da banca de defesa pela coerente avaliação e valiosa colaboração científica nesta tese.

Minha gratidão à direção e colegas da Universidade do Estado do Rio Grande do Norte, pelo apoio decisivo e pela confiança no meu trabalho. Agradeço também ao INPE pela oportunidade de aperfeiçoamento acadêmico e infra-estrutura fornecida, bem como à CAPES pelo apoio financeiro através da bolsa de doutorado.

Meu muito obrigado aos pesquisadores e servidores do INPE, por todo o apoio e incentivo. Agradeço também aos colegas de curso que muito colaboraram com sugestões e grande amizade. Minha gratidão aos colegas de sala pelo agradável cotidiano e apoio diário.

Agradeço a todos que, direta ou indiretamente, contribuíram com trabalho, companheirismo ou amizade durante este período. Que Deus esteja abençoando cada um de vocês!

RESUMO

O Instituto Nacional de Pesquisas Espaciais (INPE) possui mais de 130 Terabytes de dados de imagens que cobrem 30 anos de atividades de sensoriamento remoto. A disponibilidade deste grande volume de imagens demanda técnicas adequadas de exploração destes acervos. Entretanto, temos uma capacidade limitada para extrair informações a partir destas bases de dados, uma vez que o problema crucial em extração de informações a partir de bancos de imagens de sensoriamento remoto é detectar padrões de mudança. Esta tese apresenta uma proposta para extrair padrões de mudança a partir de imagens de sensoriamento remoto utilizando conceitos de processamento digital de imagens, mineração de dados e ecologia da paisagem. Diante da problemática sócio-ambiental advinda do rápido desflorestamento da Amazônia, este trabalho fornece e avalia recursos para auxiliar na compreensão de processos de mudança de uso do solo, bem como na tomada de decisões inerentes a este domínio. A metodologia desenvolvida foi aplicada em dados de sensoriamento remoto, através de um protótipo computacional, para identificar e analisar processos de desflorestamento em áreas amazônicas.

MINING PATTERNS OF CHANGE IN REMOTE SENSING IMAGES

ABSTRACT

The National Institute for Space Research (INPE) holds more than 130 terabytes of image data, which cover 30 years of remote sensing activities. The availability of such huge image set demands appropriate techniques to explore it. However, we have a limited capacity to extract information from these databases, once the main problem of information extraction from remote sensing images is to detect patterns of change. This thesis presents a proposal to extract patterns of change from remote sensing images through concepts of digital image processing, data mining and landscape ecology. Taking into account the social and environmental problems caused by the fast Amazon deforestation, this work supplies and evaluates resources to assist the comprehension of land use and cover change processes, as well decision making procedures related to this domain. The developed methodology was applied in remote sensing data, through a software prototype, to identify and analyze the deforestation processes in Amazon areas.

SUMÁRIO

Pág.

LISTA DE FIGURAS

LISTA DE TABELAS

CAPÍTULO 1

INTRODUÇÃO

Apesar do grande sucesso dos programas globais de sensoriamento remoto e da ampla disponibilidade de dados deste domínio, torna-se cada vez mais evidente uma "lacuna semântica" no processo de extração de informações a partir de imagens. Esta "lacuna" ocorre porque nossa capacidade de construir sofisticados satélites de observação da Terra não é compatível com nossos meios de produção de informação a partir destas fontes de dados (MacDonald, 2002). Em muitos pontos temos falhado na exploração do potencial dos dados que coletamos. Uma área onde esta "lacuna semântica" é particularmente crítica é o uso de nossos imensos arquivos de imagens de sensoriamento remoto para suprir demandas sociais, ambientais, públicas e econômicas. Atividades envolvendo detecção, análise e combate à devastação da floresta amazônica figuram como exemplo de extrema relevância desta realidade.

1.1 Motivação

O Arquivo Nacional Americano de Dados de Sensoriamento Remoto da Terra por Satélite (US National Satellite Land Remote Sensing Data Archive), gerenciado pelo Centro de Dados USGS EROS, possui 1.400 terabytes de dados de satélite coletados em 40 anos (U.S. Department of the Interior, 2005), e satélites como o Terra e o Aqua da Nasa geram um adicional de 3 terabytes de imagens todos os dias (NASA, 2005). O Instituto Nacional de Pesquisas Espaciais (INPE) possui mais de 130 terabytes de dados de imagens, cobrindo 30 anos de atividades de sensoriamento remoto, os quais estão disponíveis em um banco de dados com livre acesso on-line para pesquisadores brasileiros (INPE, 2005a).

A disponibilidade de grandes arquivos de imagens de sensoriamento remoto demanda técnicas adequadas de exploração destes dados. Atualmente, temos uma capacidade limitada para extrair informações a partir destas bases de dados. Um grande banco de dados de sensoriamento remoto é uma coleção de fotos da paisagem, que fornecem uma oportunidade única de compreender como, quando e onde mudanças ocorreram em nosso mundo. O banco de dados de imagens do INPE, por exemplo, cobre três décadas de mudança de uso do solo na floresta tropical amazônica.

Trabalhos extensivos de campo indicam que os diferentes atores envolvidos na mudança de uso do solo (pequenos agricultores, grandes fazendeiros, criadores de gado) podem ser distinguidos pelos seus diferentes padrões espaciais de uso da terra (Lambin *et al.*, 2003). Além disso, estes padrões evoluem no tempo, pois novas pequenas fazendas são criadas e grandes propriedades aumentam suas áreas plantadas às custas da floresta. O desflorestamento amazônico destaca-se pela complexidade do processo e extensão da área desmatada, a qual possui uma média de 25.000 Km^2 por ano. Diante da degradação do solo e da problemática sócio-ambiental advinda do rápido desflorestamento, quanto mais rápida e precisa a identificação dos atores e processos em áreas de desmatamento, maiores as possibilidades de conter a devastação e suas conseqüências.

Neste contexto, e em outros similares, padrões de uso da terra terão assinaturas espectrais semelhantes, e técnicas de extração de conhecimento baseadas em agrupamento no espaço amostral não serão capazes de distingui-los. Portanto, consideramos que o problema crucial em extração de informações a partir de bancos de imagens de sensoriamento remoto é *detectar padrões de mudança de uso do solo*.

1.2 Proposta da Tese

O presente trabalho propõe métodos para extração de informações semânticas a partir de imagens de sensoriamento remoto através de técnicas de mineração de imagens.

Para efetuar este processo de busca, propomos uma metodologia de mineração que utiliza técnicas de extração de informação que remetem ao domínio de dados de sensoriamento remoto: segmentação de imagens, métricas de ecologia da paisagem e classificadores baseados em árvores de decisão. A metodologia demanda interação com um ou mais especialistas do domínio minerado, uma vez que o conhecimento da aplicação e das questões relacionadas à mesma é um requisito fundamental para a obtenção de resultados satisfatórios no processo de mineração.

Nossa abordagem é motivada por trabalhos anteriores do nosso grupo de pesquisa no uso de ontologias para GIS integrados (Fonseca *et al.*, 2002), e em caracterização ontológica de imagens de sensoriamento remoto (Câmara *et al.*, 2001a). A partir da metodologia desenvolvida e dos resultados obtidos nesta pesquisa, tivemos um artigo aceito em

conferência internacional (The Fifth IEEE International Conference on Data Mining, 2005) (Silva *et al.*, 2005b) e um trabalho submetido a revista internacional (International Journal of Remote Sensing) (Silva *et al.*, 2005a). Ambos encontram-se no apêndice deste documento.

1.3 Questão Científica, Hipótese de Trabalho e Objetivos

Diante do contexto exposto, a questão científica subjacente desta tese pode ser colocada da seguinte forma: *como extrair padrões de mudança de uso do solo em imagens de sensoriamento remoto?*

A hipótese de trabalho adotada propõe que *a partir dos conceitos de processamento digital de imagens, mineração de dados e ecologia da paisagem é possível desenvolver uma metodologia de extração de padrões de mudança em imagens de sensoriamento remoto*. A tese propõe uma metodologia de mineração de padrões de mudança de uso do solo em imagens de sensoriamento remoto, com os seguintes passos:

- Extração de métricas de ecologia da paisagem que retratem objetos de imagens, permitindo obter uma caracterização estrutural do acervo analisado;

- Especificação de tipologias de padrões espaciais a partir de atributos estruturais e conceitos dos domínios de aplicação inerentes às imagens analisadas, visando caracterizar elementos da paisagem;

- Mineração dos padrões espaciais com algoritmo de aprendizagem de máquina, objetivando extrair informações semânticas que identifiquem padrões de mudança de uso do solo;

A partir da metodologia proposta, foram obtidos os seguintes resultados:

- Desenvolvimento de um protótipo computacional capaz de detectar padrões de mudança de uso do solo em imagens de sensoriamento remoto;

- Aplicação da metodologia, através do protótipo, para identificar e analisar processos de desflorestamento em áreas da Amazônia utilizando dados de sensoriamento remoto;

- Avaliação comparativa dos resultados obtidos através da metodologia com dados e estudos de campo consolidados que retratam processos de desflorestamento e seus fatores subjacentes;

1.4 Contribuição da Tese

As atuais ferramentas de mineração de dados não oferecem recursos suficientes para investigar questões relacionadas a mudanças temporais retratadas em imagens de sensoriamento remoto. Mesmo sendo uma área de pesquisa e aplicação relevante e promissora diante das demandas relacionadas aos imensos acervos de imagens, limitações que vão de representação espacial a especificação semântica impedem que muitas propostas de mineração e tratamento de imagens extraiam informações estratégicas provenientes da dinâmica da paisagem.

Este trabalho busca superar estas limitações utilizando técnicas de mineração de dados (classificador por árvore de decisão), processamento digital de imagens (segmentação por crescimento de regiões) e ecologia da paisagem (métricas de mapas categóricos). Fornecendo e avaliando recursos para auxiliar na compreensão de processos de mudança de uso do solo, bem como na tomada de decisões inerentes a este domínio, a proposta permite a transformação de imagens em informações estratégicas através de uma metodologia de mineração de imagens de sensoriamento remoto. Utilizando uma abordagem exploratória e qualitativa destes processos, a metodologia permite extrair padrões que possibilitam ao especialista (ecólogo, biólogo, geógrafo, dentre outros) compreender os processos de mudança, levando em consideração o fator temporal inerente à transformação de paisagens.

1.5 Organização da Tese

Este trabalho está organizado em cinco capítulos. O segundo capítulo aborda a mineração de imagens e traz uma revisão bibliográfica da área. Considerações sobre mineração de imagens e sobre os trabalhos citados também compõem esta parte do documento.

No terceiro capítulo descrevemos a metodologia proposta para extrair padrões de mudança a partir de imagens de sensoriamento remoto. Pontos como a definição da tipologia e a construção de conjuntos de referência de padrões espaciais são abordados nesta fase.

Explicamos como a mineração de imagens é realizada através de um classificador estrutural e de métricas de ecologia da paisagem. O protótipo implementado (PattFinder) também é descrito neste capítulo.

A aplicação da metodologia é apresentada no quarto capítulo, cujo objetivo é obter uma melhor compreensão do processo de mudança de uso do solo em áreas da Amazônia. Realizamos uma abordagem da questão amazônica e, utilizando o PattFinder, demonstramos a mineração de padrões em dados sintéticos e dados de imagens. Regiões de São Félix do Xingu (PA), do Vale do Anari (RO) e da Terra do Meio (PA) são tomadas como estudos de caso na obtenção de informações estratégicas através da mineração de imagens.

O quinto capítulo traz uma discussão sobre a metodologia proposta, seus resultados, contribuições, limitações e trabalhos futuros.

CAPÍTULO 2

REVISÃO BIBLIOGRÁFICA

Este capítulo apresenta uma breve descrição de técnicas de mineração de imagens e sua aplicação em dados de sensoriamento remoto. Mineração de imagens é um processo que agrega tanto características como demandas distintas dos métodos de mineração de dados convencionais. Em diferentes áreas e aplicações, propostas buscam extrair informações e padrões a partir de imagens através de técnicas inerentes a cada domínio minerado. Neste capítulo abordamos a mineração de imagens, suas especificidades e os trabalhos relevantes que mineram informações contidas em imagens médicas, astrofísicas e de sensoriamento remoto.

2.1 Mineração de Imagens

Descoberta de conhecimento em bancos de dados (DCBD) é o processo de identificar em dados padrões que sejam válidos, previamente desconhecidos, potencialmente úteis e compreensíveis, visando melhorar o entendimento de um problema ou um procedimento de tomada de decisão (Fayyad *et al.*, 1996). O processo de DCBD é interativo, iterativo, cognitivo e exploratório, envolvendo passos como: definição do tipo de conhecimento a descobrir, seleção de dados alvo, pré-processamento, transformação, mineração destes dados, subseqüente interpretação de padrões e implantação do conhecimento descoberto. A cada etapa, passos anteriormente realizados podem ser retomados com base em observações e descobertas realizadas até aquele momento, visando solucionar problemas encontrados ou aumentar a qualidade dos resultados. Mineração de dados é a etapa em DCBD responsável pela seleção dos métodos a serem utilizados para localizar padrões nos dados, seguida da efetiva busca por padrões de interesse numa forma particular de representação, juntamente com a busca pelo melhor ajuste dos parâmetros do algoritmo para a tarefa em questão. Mineração de dados em imagens utiliza técnicas de DCBD e de outras áreas, respeitando a complexidade e a amplitude semântica deste domínio.

O processo de mineração de dados em imagens segundo Zhang (Zhang *et al.*, 2002) é apresentado na Figura 2-1. As imagens de um acervo são recuperadas segundo critérios inerentes à aplicação. A seguir, uma fase de pré-processamento aumenta a qualidade dos

dados, os quais são então submetidos a uma série de transformações e de extração de características que geram importantes informações a respeito das imagens. A partir destas informações, a mineração pode ser realizada através de técnicas específicas, com o intuito de descobrir padrões significativos. Os padrões resultantes são então interpretados e avaliados para a obtenção do conhecimento final, que pode ser aplicado no entendimento de problemas, na tomada de decisões ou em outras atividades estratégicas.

A mineração de dados em imagens não consiste simplesmente no uso de técnicas de mineração de dados aplicadas em bancos "convencionais" ao domínio de imagens. Diferenças importantes entre estes bancos convencionais e os de imagens incluem:

- *Textura* – numa imagem, cada elemento (pixel) está muito relacionado com seus vizinhos, muitas vezes fazendo parte de uma região homogênea. Se tratarmos os elementos de imagens como entidades isoladas, perdemos a capacidade de capturar a informação de textura presente no contexto.

- *Processamento em vários níveis* – estudos clássicos sobre a visão humana (Marr, 1982) mostram que o processamento visual ocorre em vários níveis. A princípio ocorre a detecção de bordas, geometrias e estruturas dos objetos, até chegar à identificação de objetos na cena, contextualizando-os de acordo com as intenções e conhecimento do observador, associando os elementos perceptuais (borda, geometria, estrutura) a padrões, protótipos e eventos segundo a cognição do indivíduo.

- *Ambigüidade de interpretação* – diferentes observadores podem interpretar a mesma imagem de forma distinta, dependo da natureza do estudo e dos métodos de análise empregados. O nível de conhecimento e experiência do intérprete influenciam diretamente na compreensão e avaliação dos elementos retratados.

- *Dependência de domínio* – cenários e atividades do mundo real pertencentes a determinado domínio possuem características e elementos específicos. A identificação dos elementos, suas classes e relacionamentos está ligada ao contexto em si, podendo uma mesma imagem possuir informações distintas e inerentes a diferentes domínios em questão.

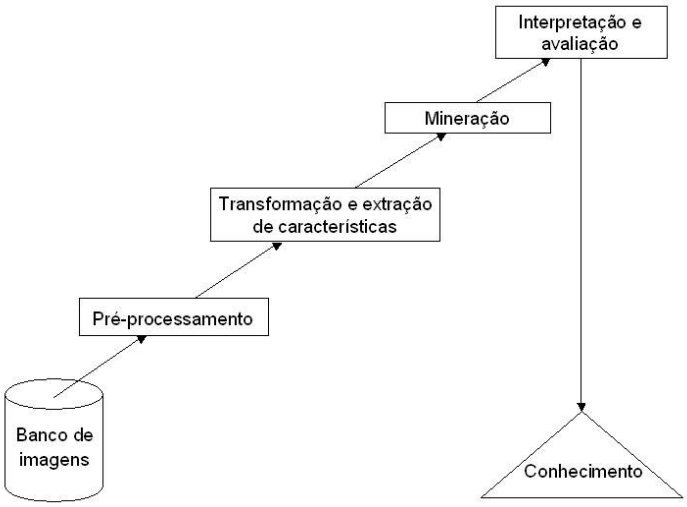


Figura 2-1 – Processo de mineração de imagens
FONTE: (Zhang *et al.*, 2002)

Devido a estas diferenças, a mineração de dados em imagens é um procedimento em várias etapas, pois assim é possível obter em cada passo requisitos e informações que serão fundamentais para a etapa subseqüente, respeitando desta forma a contextualização dos dados, a dependência do domínio e superando potenciais ambigüidades das informações.

Sendo uma área relativamente nova, com diferentes aplicações e estágios de avanço tecnológico, a mineração de imagens é implementada através de metodologias e arquiteturas que suportam variados domínios. Isto demonstra um grande interesse das comunidades usuárias de imagens (sensoriamento remoto, medicina, astrofísica, dentre outros) em avanços científicos significativos, haja vista a grande oferta de imagens e a eminente demanda de informações estratégicas oriundas destas imagens.

Cada domínio de aplicação possui necessidades e técnicas específicas de mineração, apesar de terem um conjunto de elementos e funcionalidades genérico. Relevantes domínios de aplicação de mineração de imagens são apresentados a seguir, contextualizando importantes atividades de pesquisa nesta área.

2.2 Mineração de Imagens de Sensoriamento Remoto: Propostas da Literatura

Imagens de sensoriamento remoto possuem uma natureza multiespectral, retratando formas arbitrárias do ambiente imageado que variam no tempo. Neste caso, não existem padrões fixos, mas feições de um domínio que retratam uma realidade e sua evolução. Pela riqueza de informações, e papel estratégico das imagens de sensoriamento remoto, estas têm sido alvo de diferentes iniciativas de projeto e pesquisa. São elencados a seguir alguns trabalhos relevantes de mineração de imagens de sensoriamento remoto, que abordam desafios inerentes aos seus respectivos domínios.

(Nagao; Matsuyama, 1980) desenvolveram na Universidade de Kyoto o primeiro sistema de visão de alto nível para interpretação de imagens aéreas. Os módulos de processamento do sistema operam em função de uma base de dados comum. O processo de análise no sistema é dividido nas seguintes etapas: suavização, onde imagens são suavizadas para remover ruídos e manchas nas bordas; segmentação, para extrair regiões elementares através de um algoritmo básico de crescimento de regiões sem incorporar qualquer informação adicional ao objeto; exame global da cena, para estimar domínios de objetos aproximados utilizando metadados da imagem; análise detalhada de áreas, quando subsistemas de detecção de objetos analisam uma base de conhecimento para localizar objetos específicos; comunicação entre subsistemas de detecção de objetos, que controla o fluxo da análise gerenciando a informação na base de dados, soluciona conflitos entre subsistemas de detecção e corrige erros de segmentação.

O KIM (Schröder *et al.*, 2000) é um projeto aplicado a seqüências de imagens multisensores (óticos e radares), utilizando extração de características, classificação e aquisição de conhecimento através de aprendizagem interativa, inclusive modelos bayesianos. Diferentes níveis, que vão dos dados de pixel à semântica específica do usuário, são utilizados para extrair informações e representá-las segundo a demanda colocada, tentando combinar corretamente diferentes algoritmos e modelos, além de uma adequada interação com o usuário, com vistas a uma otimização de resultados (Figura 2-2). O sistema possui recursos para acessar grandes coleções de imagens de sensoriamento remoto, através de técnicas de recuperação de imagens baseada em conteúdo e mineração de informações.

Para tarefas de detecção de alvos, por exemplo, o sistema utiliza dentre outros: parâmetros de textura (extração de informações estruturais), assinatura espectral de alvos (descrição de conteúdo da imagem), bandas não-correlacionadas geradas por Análise de Componentes Principais (segregação de componentes ruidosos e redução de dimensionalidade dos dados), distância angular espectro-temporal (obtenção de independência em relação à iluminação), Índice de Vegetação por Diferença Normalizada (distinção entre superficies com vegetação e outros elementos como nuvem, neve, água etc.). A estrutura de software e hardware, acessada através de clientes com navegadores web, é composta por sistema gerenciador de banco de dados, servidor web, servidor de mapas e imagens, banco de imagens e softwares utilitários.

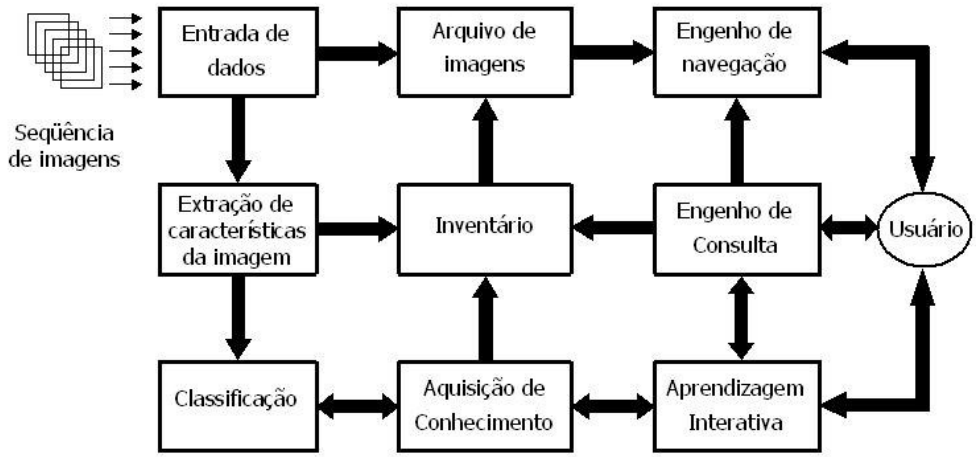


Figura 2-2 - Modelo lógico do KIM

FONTE: (Schröder *et al.*, 2000)

Redes neurais artificiais tentam alcançar o desempenho humano em vários campos, inclusive a "compreensão" de imagens. Uma rede neural é um processador paralelo massivamente distribuído, composto de unidades básicas de processamento, onde cada uma tende ao armazenamento de conhecimento experimental, tornando tal conhecimento disponível para o uso. Alguns dos pontos fortes destas redes são o reconhecimento de padrões, a previsão de tendências e a construção de modelos de dados. Um projeto relevante de mineração de imagens, baseado em redes neurais, avalia e detecta mudanças em séries temporais de imagens de sensoriamento remoto (Clifton, 2003). O trabalho apresenta uma técnica para o uso de modelagem preditiva para identificar mudanças incomuns em imagens (sem uma

noção pré-definida do que é comum ou incomum). As redes são treinadas para prever valores de regiões numa seqüência de imagens. Diferenças substanciais entre os valores esperados e os reais representam então uma mudança incomum.

O Visimine (Aksoy *et al.*, 2004) possui uma infraestrutura e metodologia para análise de imagens de satélite (Figura 2-3). O sistema trabalha com níveis de pixel, região e tiles (partições da imagem). No nível de pixel, informações espectrais e texturais sobre cada pixel são obtidas, enquanto dados poligonais descrevem a conexão entre estes pixels e características das bordas do polígono. Segmentações, histogramas e filtros são alguns dos recursos utilizados na extração de informações das imagens. As características e as imagens são indexadas e armazenadas em um sistema de banco de dados (SGBD). Dados de modelo de elevação digital (DEM) também são suportados pela arquitetura. Sistemas de informação geográfica (SIG) são utilizados para manipulação e representação de dados espaciais, ao passo que um pacote estatístico fornece recursos para análise estocástica. Funcionalidades como busca de regiões por similaridades, agrupamento, classificação, regressão, modelos bayesianos e análise de mistura espectral são disponibilizados. Uma linguagem semelhante a SQL é utilizada para consultas à base de dados, enquanto uma interface gráfica permite navegar e manipular imagens e dados associados.

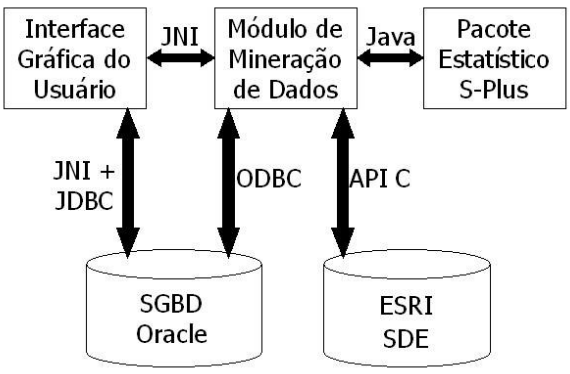


Figura 2-3 – Arquitetura do Visimine

FONTE: (Aksoy *et al.*, 2004)

ADaM, projeto da NASA em conjunto com a Universidade de Alabama em Huntsville, é um conjunto de ferramentas de mineração de dados científicos e de imagens (Rushing *et al.*, 2005). Suas funcionalidades incluem reconhecimento de padrões, processamento de

imagens, otimização, mineração de regras de associação, dentre outros. O sistema é composto por uma série de componentes individuais que podem ser utilizados em conjunto para realizar tarefas complexas. O software possui módulos implementados em C, C++ e componentes Python, com programas executáveis em linha de comando. Um dos focos do projeto é a implementação eficiente de componentes de desempenho crítico, além do cuidado de manter cada componente do sistema o mais independente possível, visando possibilitar a utilização de subconjuntos de módulos apropriados para determinadas aplicações, inclusive aproveitando componentes de terceiros.

2.3 Mineraçào de Imagens de Sensoriamento Remoto: O problema da busca de padrões de mudança

Os trabalhos de recuperação, tratamento e mineração de imagens apresentados neste capítulo abordam aspectos distintos e relevantes na obtenção de padrões e conhecimento nas imagens de sensoriamento. No entanto, estas técnicas têm limitações quando usadas para detecção de padrões de mudança em imagens de sensoriamento remoto. A detecção de padrões de mudança de uso do solo, por exemplo, permite compreender como, onde, quando e quem promove alterações em florestas e regiões estratégicas do planeta.

Para detectar padrões de mudança, é necessário compreender as diferenças entre os dados de sensoriamento remoto e outros tipos de imagens. Imagens adquiridas por sensores remotos são recursos capazes de capturar a dinâmica da paisagem. Uma paisagem geográfica é um cenário em mudança contínua, e o processo de aquisição de dados através de satélites de sensoriamento remoto origina imagens que são medições capazes de capturar momentos de trajetórias de mudança. O desafio para técnicas de mineração de imagens neste domínio consiste em descrever um processo contínuo baseado em dados que retratam alguns momentos. Desta forma, a ênfase não deve ser dada a procedimentos de detecção e identificação de objetos, mas à captura da dinâmica em uma paisagem finita.

Capturar a evolução temporal de padrões em imagens de sensoriamento remoto requer métodos distintos daqueles utilizados em sistemas de recuperação de imagem baseados em conteúdo (content-based image retrieval systems - CBIR). Estes sistemas geralmente utilizam uma imagem de referência, e imagens no banco de dados como alvos da mineração, ao passo que o resultado do processo é um conjunto de imagens classificado pela

similaridade de características com a imagem de referência (Chen *et al.*, 2003). Para detectar padrões em acervos de imagens de sensoriamento remoto, uma abordagem diferente é necessária: ao invés de buscas por similaridade entre pares de imagens, um sistema de mineração de imagens de sensoriamento remoto deve ser capaz de extrair e descrever padrões encontrados em diferentes imagens. Portanto, mineração de imagens de sensoriamento remoto é *a busca por padrões de mudança e não por conteúdos internos*. Estes conceitos serão explorados na metodologia proposta no Capítulo 3.

CAPÍTULO 3

METODOLOGIA DE MINERAÇÃO DE PADRÕES DE MUDANÇA EM IMAGENS DE SENSORIAMENTO REMOTO

Neste capítulo descrevemos a metodologia proposta para extrair padrões de mudança a partir de imagens de sensoriamento remoto. Tomamos, para isto, a motivação proveniente da disponibilidade de imensos repositórios de imagens de sensoriamento remoto. Pesquisadores gostariam de explorar estes dados com questões como: *Quais são os diferentes padrões de uso do solo presentes no banco de dados? Quando certo padrão de uso do solo surgiu? Quais são os padrões de uso do solo dominantes para cada região? Como padrões surgem e mudam ao longo do tempo?* As respostas a estas e outras questões semelhantes requerem a disponibilidade de técnicas de mineração de dados que sejam capazes de realizar buscas de similaridade de padrões de mudança em diferentes imagens. Propomos a abordagem deste problema através do uso de padrões espaciais, os quais suportarão a descrição de características semânticas relevantes de uma imagem.

3.1 O Processo de Mineração de Imagens

Uma visão geral do processo de mineração de imagens de sensoriamento remoto, segundo a nossa metodologia, é apresentada na Figura 3-1. Num primeiro passo, imagens são *selecionadas* a partir de um repositório de acordo com as necessidades da aplicação. Uma fase de pré-processamento, consistindo de *calibrações geométricas e radiométricas* aumentam a qualidade dos dados. A seguir, as imagens passam por um procedimento de segmentação e classificação de regiões, cujos resultados são regiões identificáveis (rotuladas) na imagem com bordas bem definidas.

A fase de *mineração* consiste da atribuição de descrições a estas regiões, e à identificação de quais delas pertencem a uma dinâmica da paisagem específica (conjunto de objetos, seus relacionamentos e agentes que interagem naquela área). Por exemplo, uma busca em imagens por áreas desflorestadas devido à criação de gado requer a princípio um procedimento de segmentação que distinga e rotule cada região da imagem. A seguir, um algoritmo de mineração deve ser treinado para reconhecer as regiões correspondentes a áreas

de criação de gado, com base na "assinatura espacial" (características específicas das áreas desflorestadas) destes agentes (criadores de gado). Com base nesta *mineração da dinâmica*, os *padrões de mudança* são identificados de acordo com requisitos espaço-temporais. Por exemplo, o processo pode identificar um padrão de mudança que representa um aumento nas atividades de criação de gado durantes os últimos cinco anos numa área específica.

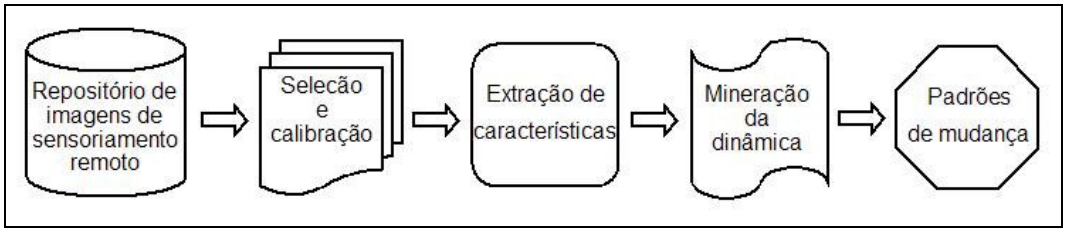


Figura 3-1 – Visão geral da metodologia de mineração de padrões

3.2 Visão Geral da Metodologia

Esta metodologia considera que instrumentos a bordo de satélites de sensoriamento remoto capturam energia em diferentes segmentos do espectro eletromagnético, a qual é então convertida em imagem digital. Estes instrumentos não são projetados para uma aplicação específica, mas são um compromisso entre a tecnologia de sensoriamento e requisitos de diferentes comunidades de usuários. Como resultado, imagens de sensoriamento remoto possuem uma descrição estrutural que independe do domínio de aplicação que um cientista utiliza para extrair informação. Portanto, precisamos distinguir entre o domínio da imagem e o domínio da aplicação, conforme apresentado na Figura 3-2:

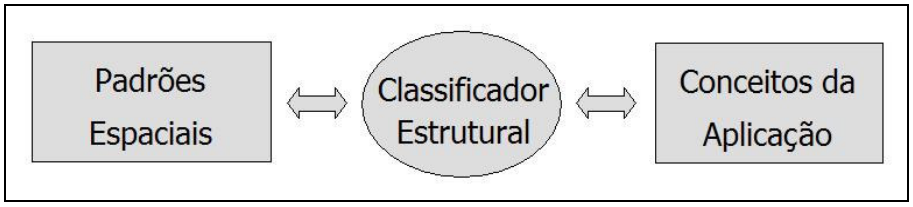


Figura 3-2 – Visão geral da classificação estrutural

- *Padrões Espaciais* – as estruturas geométricas que podem ser obtidas a partir de imagens utilizando técnicas de extração de características, como segmentação e classificação de imagens. Estas devem ser identificadas e rotuladas de acordo com uma tipologia que expresse sua semântica. Exemplos de tais padrões incluem regiões

com formato de corredor e com formato de polígonos regulares representando tipos de padrões dos dados minerados.

- *Conceitos da Aplicação* – as diferentes classes de objetos espaciais que são associadas a um domínio específico. Por exemplo, em avaliações de desflorestamento, conceitos incluem agricultura de larga escala, agricultura familiar, criação de gado e extração de madeira.

Para associar estruturas encontradas na imagem a conceitos da aplicação precisamos de um *classificador estrutural*, que é capaz de relacionar as mesmas estruturas a diferentes domínios de aplicação. Esta estratégia difere da maioria dos sistemas de mineração de imagens de sensoriamento remoto, tais como KIM (Schröder *et al.*, 2000) e VISIMINE (Aksoy *et al.*, 2004), as quais implicitamente assumem que existe um melhor ajuste (*best fit*) para associar conceitos semânticos no domínio do usuário a estruturas derivadas de imagens. Em nossa visão, para cada tipo de análise de padrões de mudança, existirão diferentes associações entre padrões espaciais e os conceitos do domínio do usuário. Cada associação é válida dentro de um dado contexto de aplicação. Para cada tipo de aplicação, haverá uma tipologia de padrões de mudança e um modelo de classificação estrutural apropriados.

Nossa metodologia possui três etapas:

- Definição de uma *tipologia de padrões espaciais* de acordo com o domínio de aplicação do usuário.

- Construção de um *conjunto de referência de padrões espaciais*, utilizando imagens prototípicas.

- Mineração do banco de dados utilizando um *classificador estrutural* (orientado aos conceitos de aplicação do domínio), associando tipos representados pelo conjunto de referência de padrões espaciais aos objetos da paisagem de imagens.

3.3 Definindo uma Tipologia de Padrões Espaciais

A primeira fase da metodologia demanda a definição de uma tipologia de padrões espaciais, a qual é associada a um determinado domínio de aplicação. Para ilustrar nossa proposta,

apresentaremos tipologias definidas para mapear diferentes tipos de mudança de uso do solo em florestas tropicais.

Imagens de sensoriamento remoto são freqüentemente utilizadas para compreender os fatores que determinam a mudança de uso do solo em florestas tropicais. A suposição é de que as mudanças de uso do solo podem ser capturadas pelas propriedades espectrais e espaciais das imagens (Alves *et al.*, 2003). Extensivos trabalhos de campo também indicam que os diferentes atores envolvidos em mudança de uso do solo (pequenos agricultores, grandes fazendeiros, criadores de gado) podem ser distinguidos por seus diferentes padrões de uso do solo (Lambin *et al.*, 2003). Lambin, Geist *et al.* (2003) propuseram uma tipologia dos padrões de uso do solo segundo processos de desflorestamento em florestas tropicais, numa escala mundial através de modelos visuais (Figura 3-3), por exemplo: *corredor* (geralmente associado à colonização ao longo de estradas e rios), *difuso* (comumente relacionado à agricultura de subsistência), *espinha de peixe* (típica de esquemas de assentamento planejado) e *geométrico* (frequentemente ligado a desflorestamentos em larga escala para atividades de setores modernos).

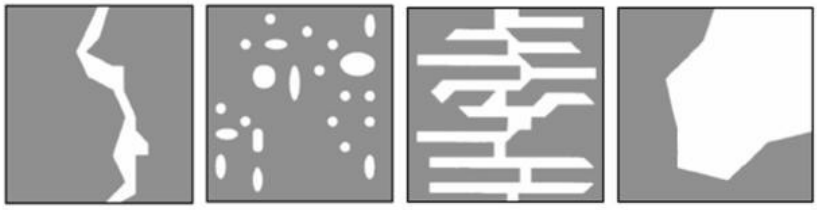


Figura 3-3 – Tipologia de padrões espaciais de desflorestamento tropical (da esquerda para a direita): corredor, difuso, espinha de peixe, geométrico (desconsiderar escala)
FONTE: (Lambin *et al.*, 2003)

Para a análise de processos em uma escala regional, precisamos definir padrões espaciais com maiores detalhes. No caso da floresta amazônica, com base em (Escada, 2003), tipologias regionais são propostas de acordo com os atores do processo de desflorestamento, dimensões das propriedades, uso do solo e padrões de desflorestamento para duas regiões: o assentamento do INCRA (Instituto Nacional de Colonização e Reforma Agrária) do Vale do Anari no estado de Rondônia, e a região denominada "Terra do Meio", no estado do Pará. Em Rondônia, por exemplo, o INCRA implantou diferentes estruturas de lotes de colonos em esquemas planejados de assentamentos, com base em diferentes arranjos espaciais. Como

resultado deste processo, padrões espaciais regionais são observados: irregulares pequenos, lineares, grandes geométricos (Figura 3-4).

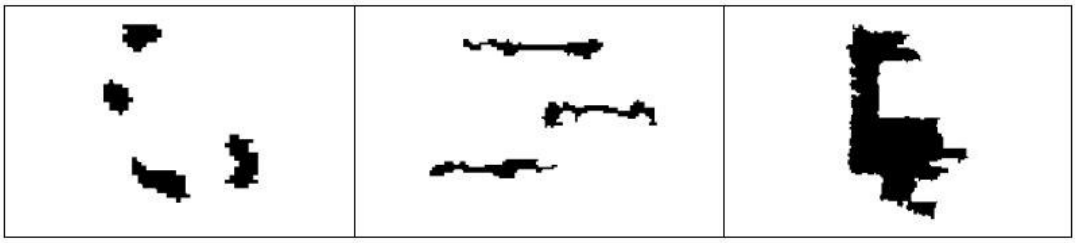


Figura 3-4 - Tipologia de padrões espaciais do Vale do Anari (da esquerda para a direita): irregulares pequenos, lineares, grandes geométricos (desconsiderar escala)

3.4 Construindo um Conjunto de Referência de Padrões Espaciais

Para representar as estruturas detectadas em imagens de sensoriamento remoto, introduzimos o conceito de *objetos da paisagem*. Um objeto da paisagem é uma estrutura detectada numa imagem de sensoriamento remoto através de um algoritmo de segmentação de imagens. Objetos da paisagem detectados podem ser associados a elementos de uma tipologia de padrões espaciais, com o intuito de caracterizar e fornecer semântica a estes objetos, os quais passam a ser denominados *padrões espaciais* (Figura 3-5). Para construir um conjunto de referência de *padrões espaciais*, devemos obter um conjunto de objetos prototípicos da paisagem, os quais são extraídos de imagens amostrais. Através de um processo de *avaliação cognitiva*, um especialista do domínio (um ecólogo, por exemplo) analisa estes objetos prototípicos da paisagem e os associa a elementos da *tipologia de padrões espaciais* utilizada, resultando em um conjunto de referência de *padrões espaciais* para a área em estudo (Figura 3-6).

Utilizamos algoritmos de segmentação para particionar a imagem em regiões que são espacialmente contínuas, disjuntas e homogêneas. Recentes levantamentos (Meinel; Neubert, 2004) indicam que abordagens de crescimento de regiões (Zucker, 1976) são apropriadas para produzir regiões fechadas e homogêneas. Em nossa proposta, adotamos o algoritmo de segmentação por crescimento de regiões desenvolvido pelo INPE (Bins *et al.*, 1996), e incluído no sistema SPRING (Sistema de Processamento de Informações Georeferenciadas) (Câmara *et al.*, 1996). Este algoritmo tem sido extensivamente validado para extrair padrões

de uso do solo em florestas tropicais (Shimabukuro *et al.*, 1998) e tem sido muito bem avaliado em recente levantamento (Meinel *et al.*, 2004).

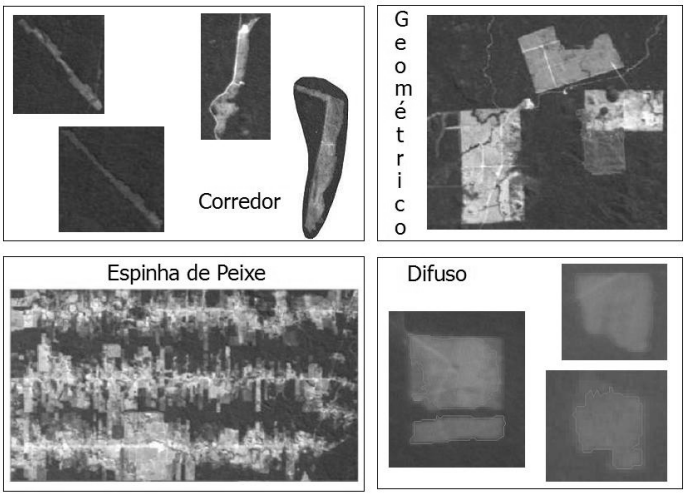


Figura 3-5 - Conjunto de referência de padrões espaciais utilizando a tipologia de (Lambin *et al.*, 2003) (desconsiderar escala)

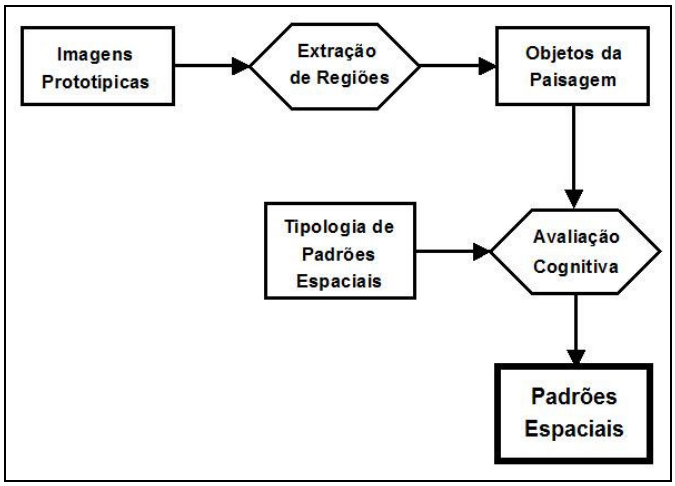


Figura 3-6 - Construindo um conjunto de referência de padrões espaciais

O algoritmo de segmentação por crescimento de regiões utilizado funciona da seguinte forma (Bins *et al.*, 1996):

- A imagem é primeiramente segmentada em células atômicas de um ou poucos pixels;

- Cada segmento é comparado com seus vizinhos para determinar se eles são similares ou não. Se forem similares, eles são unificados e o nível de cinza médio do novo segmento é atualizado;

- O segmento continua crescendo através da comparação com todos os seus vizinhos, até que não haja mais regiões passíveis de unificação, quando então o segmento é rotulado como uma região completa;

- O processo desloca-se para a próxima célula incompleta, repetindo toda a seqüência, até que todas as células estejam rotuladas.

O algoritmo requer dois parâmetros: (a) um valor de limiar de similaridade, e (b) um valor de limiar de área. Um exemplo de segmentação de imagens é apresentado na Figura 3-7.

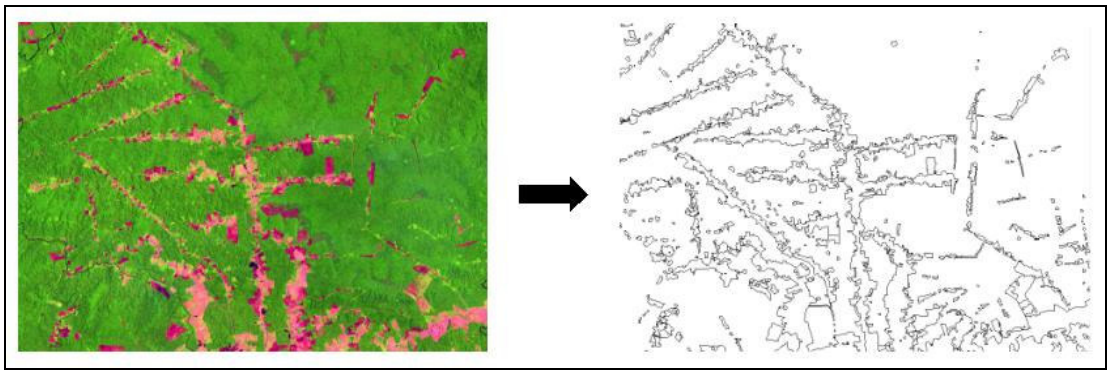


Figura 3-7 - Exemplo de segmentação de imagens

3.5 Minerando o Banco de Dados através de um Classificador Estrutural

Uma vez construído o conjunto de referência de *padrões espaciais*, nesta próxima fase ele será utilizado para minerar *configurações espaciais*. Em nossa metodologia, um *objeto da paisagem* é uma região da imagem detectada por um segmentador. Uma *configuração espacial* é um conjunto de objetos de paisagem que pertence ao mesmo tipo de padrão de mudança. Assim, a configuração espacial do tipo 'geométrico' corresponde a todas as regiões das imagens identificadas como sendo deste tipo. O *classificador estrutural* possibilita associar *objetos da paisagem* extraídos de imagens a tipos representados por elementos do conjunto de referência de *padrões espaciais* (Figura 3-8).

O *classificador estrutural* deve distinguir *objetos da paisagem* segundo os *padrões espaciais* de referência. Ele utiliza o classificador C4.5 (Quinlan, 1993), um método de classificação baseado em uma árvore de decisão (Figura 3-9). Ele prevê o valor de um atributo categórico baseado em atributos não-categóricos: o atributo categórico é o tipo do padrão - na Figura 3-9 representado por *IRR* (irregular), *GEO* (geométrico) e *LIN* (linear); os atributos não-categóricos são um conjunto de atributos numéricos, os quais são representados na Figura 3-9 pelas métricas de ecologia da paisagem que caracterizam cada tipo de padrão – *FRAC* (índice de dimensão fractal) e *SHAPE* (índice de formato da região). As idéias básicas do C4.5 são:

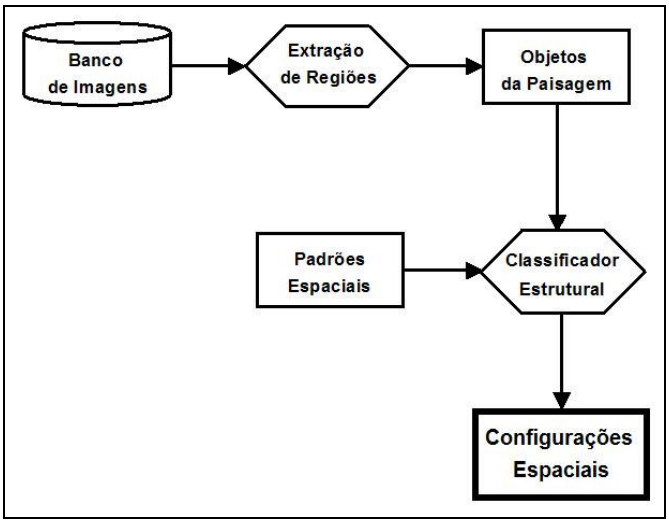


Figura 3-8 - Obtendo configurações espaciais

- Na árvore de decisão, cada nó corresponde a um atributo não-categórico, e cada arco a um possível valor daquele atributo. Uma folha da árvore especifica o valor esperado do atributo categórico para os registros descritos pelo caminho entre a raiz e esta folha;

- Na árvore de decisão, em cada nó deve ser associado o atributo não-categórico que é mais informativo entre os atributos ainda não considerados no caminho a partir da raiz;

- Entropia, uma medida da Teoria da Informação, é utilizada para medir quão informativo é um nó. Quanto maior a entropia, mais informação será necessária para

descrever os dados. O objetivo é atribuir ao nó o atributo que minimize a entropia dos dados.

Outros algoritmos de mineração foram avaliados para serem usados nesta metodologia: PART, OneR, Prism, Id3, NaiveBayes, Decision Table e Redes Neurais. A escolha do classificador C4.5 por árvore de decisão como a melhor alternativa merece algumas considerações. Em aplicações típicas de detecção de padrões de mudança em imagens de sensoriamento remoto, o número de elementos da tipologia é reduzido (tipicamente, entre três e onze tipos de padrão). Além disso, o número de amostras de treinamento é também limitado. Nestas condições, o C4.5 obteve os melhores resultados com conjuntos de dados classificados por cada um dos algoritmos citados. Devido a restrições de tipo de dados (não-categóricos), o Prism e o Id3 não conseguiram gerar o modelo.

Além de ter um desempenho superior aos demais classificadores testados, o algoritmo de classificação por árvore de decisão C4.5 foi escolhido para esta metodologia devido aos seguintes fatores: (a) o algoritmo suporta atributos (não-categóricos) contínuos nos nós da árvore, no nosso caso, métricas de ecologia da paisagem; (b) o C4.5 é largamente utilizado, testado e validado, o que indica a sua qualidade enquanto método computacional; (c) o algoritmo tenta manter a árvore a menor possível, uma vez que árvores menores são mais facilmente compreendidas e têm bom desempenho preditivo (Kohavi; Quinlan, 2002).

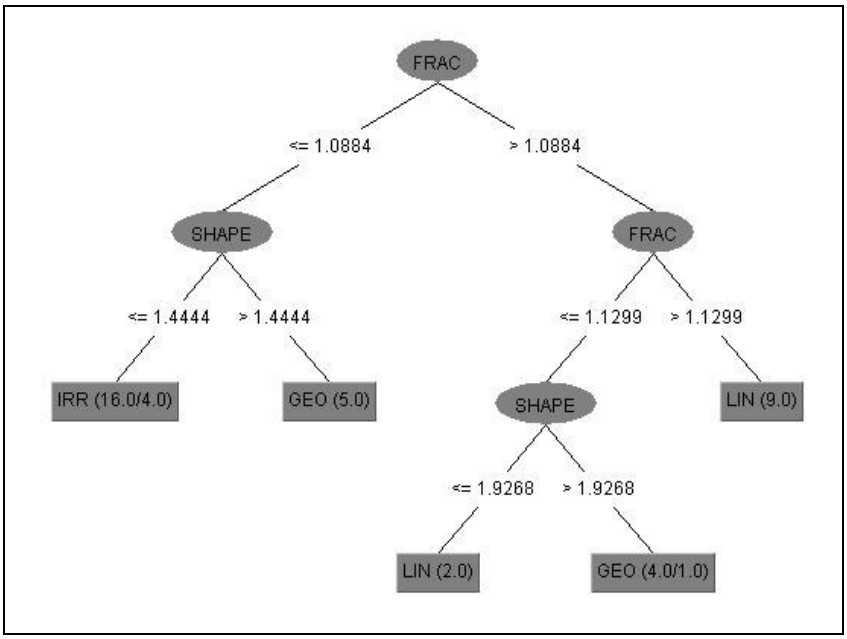


Figura 3-9 - Exemplo de árvore de decisão

Os valores que figuram ao lado de cada rótulo nas folhas da árvore de decisão indicam o número de instâncias do treinamento cujos atributos não-categóricos levam àquela folha. Quando um segundo valor aparece ao lado, isto indica que o rótulo desta instância difere do rótulo da folha, ou seja, classificação incorreta. Por exemplo: *IRR (16.0/4.0)* indica que os atributos não-categóricos de 16 instâncias levam a esta folha, e que 4 deles possuem rótulos diferentes de *IRR*, ou seja, de 16 instâncias classificadas 4 estão incorretas.

3.6 Métricas de Ecologia da Paisagem

Para selecionar os atributos que distinguem os diferentes tipos de padrão de uso do solo, utilizamos os conceitos de ecologia da paisagem (Turner, 1989). Ecologia da paisagem é baseada na noção de que padrões ambientais influenciam fortemente processos ecológicos. Um componente chave da teoria de ecologia da paisagem é a definição de *métricas* que caracterizam propriedades geométricas e espaciais de padrões de mapas categóricos (McGarigal, 2002). Algumas das métricas de padrões utilizadas em ecologia da paisagem operam no nível de regiões que, no nosso caso, são manchas ou clareiras abertas numa floresta. Regiões são os blocos elementares para mapas categóricos e a heterogeneidade intra-região é ignorada. Métricas de regiões referem-se a características espaciais como área, perímetro, complexidade de forma, compactação, contigüidade, circularidade ou convolução destas regiões na paisagem. Utilizamos as seguintes métricas implementadas pelo software FRAGSTATS (Programa de Análise de Padrões Espaciais para Mapas Categóricos) (McGarigal; Marks, 1995):

- *Perímetro* (metros):

  $PERIM = p_{ij}$, onde i é o identificador da paisagem (mapa categórico), j é o identificador da região propriamente dita, e p_{ij} é o perímetro (metros) da região i na paisagem j

- *Área* (hectares):

  $AREA = a_{ij}\left(\frac{1}{10000}\right)$, onde a_{ij} é a área (metros quadrados) da região i na paisagem j

- *Para*, razão perímetro-área, uma medida simples de complexidade do formato da região:

  $PARA = \frac{p_{ij}}{a_{ij}}$, onde p_{ij} é o perímetro (m) e a_{ij} é a área (m^2) da região *i* na paisagem *j*

- *Shape*, índice de formato da região, igual a 1 quando a região é a mais compacta possível (quadrado ou semi-quadrado), e cresce sem limite juntamente com a irregularidade do formato da região:

  $SHAPE = \frac{p_{ij}}{\min p_{ij}}$, onde p_{ij} é o perímetro (m) e *min* p_{ij} é o mínimo perímetro possível para uma região maximamente compacta (formato de quadrado) da correspondente área

- *Frac*, índice de dimensão fractal, o qual se aproxima de 1 para regiões com formatos muito simples como quadrados, e de 2 para regiões com muita convolução:

  $FRAC = \frac{2\ln(0{,}25\, p_{ij})}{\ln a_{ij}}$, onde p_{ij} é o perímetro (m) e a_{ij} é a área (m^2) da região *i* na paisagem *j*

- *Circle*, círculo circunscrito relacionado, igual a 0 para regiões circulares e aproxima-se de 1 para regiões que são lineares (estreitas e alongadas):

  $CIRCLE = 1 - \left[\frac{a_{ij}}{a_{ij}^s} \right]$, onde a_{ij} é a área (m^2) da região *ij* e a_{ij}^s é a área (m^2) do menor círculo circunscrito da região *i* na paisagem *j*

- *Contig*, índice de contigüidade, aproxima-se de 0 para regiões com apenas um pixel e de 1 conforme o aumento da contigüidade ou conectividade da região:

$$CONTIG = \frac{\left[\frac{\sum_{r=1}^{z} c_{ijr}}{a_{ij}}\right] - 1}{v - 1}$$, onde c_{ijr} é o valor de contigüidade para o pixel *r* na região *ij*, *v* é a soma dos valores numa grade de pixels 3x3 e a_{ij} é a área da região *ij*

- Gyrate, raio do giro, afetado pela extensão e compactação da região, atingindo seu valor máximo quando a região cobre toda a paisagem:

  $$GYRATE = \sum_{r=1}^{z} \frac{h_{ijr}}{Z}$$, onde h_{ijr} é a distância (m) entre o pixel *ijr* e o centróide da região *ij* e Z é o número de pixels na região *ij*

Visando manter uma uniformidade de notação, mantivemos os índices *ij*, embora na prática a extração de métricas sempre foi realizada em um mapa categórico por vez.

As métricas de ecologia da paisagem são processadas pelo *classificador estrutural* através do algoritmo de classificação C4.5 para distinguir os diferentes tipos de padrões espaciais. Após o devido treinamento deste classificador, ele pode ser utilizado para rotular os objetos da paisagem encontrados em outras imagens. Portanto, para cada imagem no banco de dados, este procedimento identifica o tipo de padrão espacial de cada objeto da paisagem. *Configuração espacial*, neste contexto, é um conjunto específico de padrões espaciais encontrados numa imagem, conforme exemplo da Figura 3-10.

Através da identificação de *configurações espaciais* em diferentes imagens, o usuário será capaz de avaliar o surgimento e evolução de diferentes tipos de mudança de uso do solo. Cada padrão espacial está associado a um diferente tipo de mudança. Portanto, a comparação entre configurações espaciais de imagens em diferentes localidades, e entre configurações espaciais de imagens na mesma localidade em diferentes momentos, permitirá novas percepções dos processos e atores que causam mudanças no uso do solo.

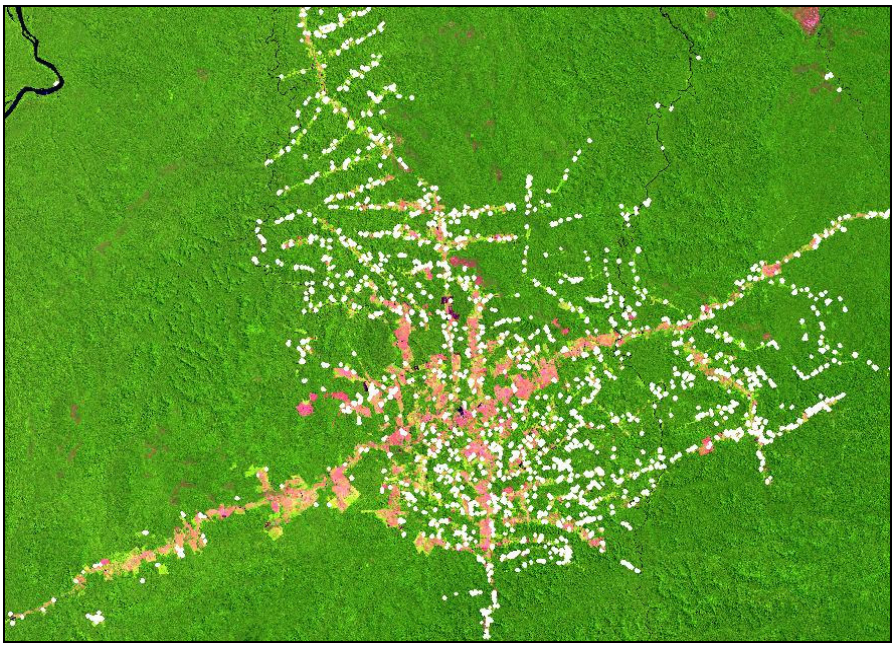


Figura 3-10 – Configuração espacial de padrões *irregulares* - Apuí (AM) – 1997 a 2003 (pontos em branco)

3.7 Protótipo Implementado - PattFinder

Diante da natural demanda advinda do desenvolvimento, teste e validação da metodologia proposta, implementamos um protótipo denominado PattFinder (do inglês *Pattern Finder* – Buscador de Padrões). No protótipo foram utilizados componentes de software que implementam funções requeridas pela metodologia: tratamento de imagens (especialmente segmentação e classificação), geração de modelo de árvore de decisão (algoritmo C4.5) e extração de métricas de ecologia da paisagem das regiões desmatadas (área, compactação, contigüidade, circularidade, dentre outros).

O protótipo, desenvolvido em C++, interage com os seguintes softwares:

- SPRING - Sistema de Processamento de Informações Georeferenciadas (Câmara *et al.*, 1996) – implementa, dentre outros, o algoritmo de segmentação por crescimento de regiões desenvolvido pelo INPE (Bins *et al.*, 1996). O software está disponível gratuitamente na internet (http://www.dpi.inpe.br/spring/).

- FRAGSTATS - Programa de Análise de Padrões Espaciais para Mapas Categóricos (McGarigal *et al.*, 1995) – extrai métricas de ecologia da paisagem de regiões de

imagens segmentadas e rotuladas (perímetro, área, índice de dimensão fractal, índice de contigüidade, dentre outros). O software encontra-se disponível na internet (http://www.umass.edu/landeco/research/fragstats/fragstats.html).

- WEKA – Waikato Environment for Knowledge Analysis (Witten; Frank, 1999) – possui uma série de algoritmos de aprendizagem de máquina implementados, inclusive o classificador por árvore de decisão C4.5 (Quinlan, 1993). O software possui código fonte aberto (Java), o qual pode ser obtido via internet (http://www.cs.waikato.ac.nz/ml/weka/).

O PattFinder integra, através das diferentes tarefas realizadas, funcionalidades dos softwares acima citados, implementando ainda funções demandadas em diferentes fases do processo. A Tabela 3-1 fornece uma visão geral das tarefas desempenhadas, funcionalidades requeridas e componentes de software que as implementam. A Figura 3-11 sintetiza o fluxo de dados entre os componentes utilizados no protótipo.

Tabela 3-1 – Tabela de funcionalidades dos softwares que compõem o protótipo

Tarefa	Funcionalidades	Software
Tratamento de imagens	Extração de regiões (segmentação/rotulação)	SPRING
	Classificação e visualização	
Avaliação cognitiva	Associação de tipos de padrões espaciais a métricas dos respectivos objetos da paisagem	PattFinder
Classificação estrutural	Extração de métricas de regiões	FRAGSTATS
	Acumulação de dados para treinamento do modelo	PattFinder
	Treinamento da árvore de decisão	WEKA
	Teste/validação do modelo	
	Aplicação do modelo às imagens	PattFinder
Integração de componentes	Sincronização de funcionalidades e fluxo de dados	PattFinder

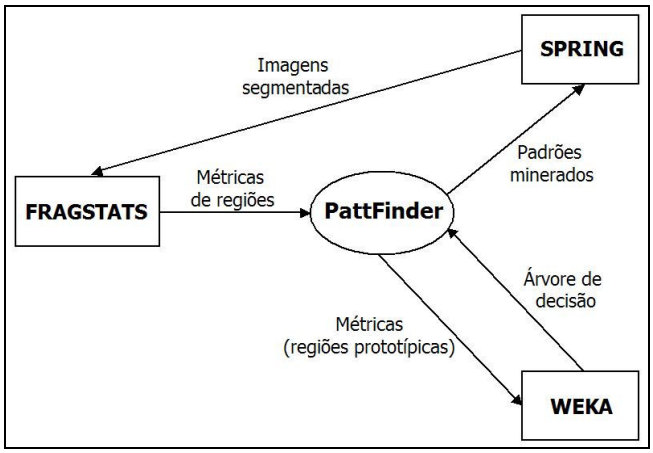


Figura 3-11 – Fluxo de dados dos componentes de software

Neste ponto, apresentamos uma breve descrição deste fluxo de dados. As imagens são visualizadas e segmentadas no Spring. Após a extração de regiões, as imagens segmentadas são exportadas para o Fragstats no formato ASCII, o qual extrai as métricas de cada região da imagem. Se estas regiões forem as prototípicas, o PattFinder auxilia no processo de avaliação cognitiva, que associará elementos da tipologia definida a objetos da paisagem e suas métricas. Estas métricas serão utilizadas pelo classificador estrutural através do algoritmo C4.5 implementado no Weka, invocado pelo PattFinder, para gerar interativamente modelos de mineração. Após a avaliação e escolha do melhor modelo (árvore de decisão), este é incorporado ao PattFinder. A partir deste momento, novas imagens serão segmentadas, exportadas para o Fragstats e terão suas métricas calculadas. O PattFinder utiliza então o modelo gerado na fase anterior para minerar padrões nas imagens através das métricas extraídas. Novas imagens retratando os padrões minerados são geradas pelo PattFinder e importadas pelo Spring, onde estes padrões poderão ser visualizados e analisados. Arquivos com dados quantitativos também são gerados e armazenados para a construção de gráficos durante o processo de análise.

CAPÍTULO 4

MINERAÇÃO DE PADRÕES DE MUDANÇA EM REGIÕES DA AMAZÔNIA

O Brasil enfrenta um grande desafio: controlar o desmatamento na floresta Amazônica, a qual cobre cerca de 40% do seu território. O desflorestamento é causado por fatores econômicos, sociais e políticos, cujo atual ritmo de mudança de uso do solo desmata anualmente uma média de 25.000 km^2 de floresta (INPE, 2005b). Tal cenário de desflorestamento demanda ações rápidas e efetivas para reduzir este ritmo de devastação, o que exige informações precisas em tempo hábil. Com o objetivo de monitorar este processo extremamente rápido de mudança de uso de solo na Amazônia, é muito importante que o INPE seja capaz de explorar ao máximo seus imensos arquivos de dados. Diante deste contexto, neste capítulo utilizamos a metodologia proposta para obter uma melhor compreensão do processo de mudança de uso do solo em áreas da Amazônia.

4.1 A Questão Amazônica

O "uso do solo", influenciado pela ação humana e pelos processos e características ambientais, está relacionado ao propósito ao qual este serve, podendo ser agricultura, habitação, extrativismo, lazer, etc. Mudanças no uso do solo ocorrem em vários níveis espaciais e em diferentes períodos, denotando a dinâmica humana e ambiental sobre segmentos territoriais. A "cobertura do solo" descreve o estado físico da superfície deste solo, podendo ser floresta, água, dentre outros. Alterações desta cobertura podem ser causadas por variações climáticas, mudanças de cursos de rios, dentre outros. Entretanto, a maioria das modificações na cobertura do solo é atribuída à ação humana. Mudança de uso e cobertura do solo significa mudanças na extensão (área – aumento ou diminuição) de um determinado tipo de uso ou cobertura (Briassoulis, 2000).

As causas e conseqüências da mudança de uso e cobertura do solo, e os impactos ambientais e sócio-econômicos desta mudança, têm motivado vários temas e projetos de pesquisa. Uma destas iniciativas é (Geist; Lambin, 2001), cujo trabalho enfatiza que mudança de uso do solo é um fator gerador de alterações globais, interagindo com clima, processos de ecossistemas, ciclos bioquímicos, biodiversidade e – ainda mais importante – com atividades humanas. A mudança de áreas de floresta para pecuária, por exemplo, é uma importante

alteração de uso do solo devido às sérias implicações envolvidas. Conseqüências como desertificação, mudança de clima e perda de biodiversidade geralmente trazem severos danos ao meio-ambiente e ao homem.

O caso amazônico é caracterizado pela complexidade, dimensão e volume de interesses envolvidos nas questões ligadas à mudança de uso e cobertura do seu solo (Becker, 1997). O trabalho de (Alves, 2002) apresenta uma investigação da dinâmica espaço-temporal do desflorestamento na Amazônia, utilizando imagens de satélite de sensoriamento remoto para a análise de padrões espaciais de desflorestamento na década de 70 e entre 1991 e 1997. O trabalho aponta que a área desflorestada passou de 10 milhões de hectares (década de 70) para aproximadamente 59 milhões de hectares em 2000; houve intensificação do desflorestamento nos anos 70 e 80 devido à política do governo federal que incluía a construção de grandes redes rodoviárias, e a apropriação de uma área de 100 Km ao longo das maiores rodovias para projetos de colonização; através da análise de imagens e seus padrões, é verificado que além do desflorestamento que se concentra ao longo das grandes rodovias e algumas zonas de desenvolvimento, ocorre ainda a fusão de pequenas áreas de desflorestamento que formam grandes áreas desflorestadas.

Uma vez que o rápido desflorestamento ocasiona degradação do solo, tensão social e urbanização precária, quanto mais rápida e precisa a identificação de áreas com tal tendência e quanto melhor a compreensão dos atores e processos envolvidos, maiores as chances de prevenir, administrar e inibir o processo e reduzir suas conseqüências. Diariamente, diferentes satélites registram dados pertinentes a este contexto, cujas imagens são disponibilizadas para diversas instituições. A metodologia proposta nesta tese pode aumentar consideravelmente a capacidade de análise deste grande acervo de informações tão estratégicas, especialmente diante das condições reunidas pelo INPE, o qual (a) conhece e pesquisa amplamente os padrões estruturais no contexto amazônico, (b) acompanha o desenvolvimento do processo histórico da Amazônia, (c) possui um rico acervo de imagens de sensoriamento remoto que fornece uma ampla cobertura espacial e temporal do território amazônico - 130 terabytes de dados cobrindo 30 anos de atividades, e (d) tem experiência no processamento e análise de imagens, bem como na produção de ferramentas de software.

4.2 Minerando Padrões em Dados Sintéticos

Os primeiros experimentos com a metodologia proposta foram realizados com dados sintéticos, com o intuito de avaliar pontualmente sua eficácia na detecção de padrões através de métricas de ecologia da paisagem. Tal eficácia só pode ser atingida se o classificador estrutural for capaz de criar modelos (árvores de decisão) capazes de abstrair satisfatoriamente os padrões através de métricas. Dado um conjunto de feições representando diferentes classes, o modelo deve distinguir e classificar cada feição de acordo com suas características estruturais retratadas pelas métricas de ecologia da paisagem (*AREA, PERIM, GYRATE, PARA, SHAPE, FRAC, CIRCLE, CONTIG*).

A Figura 4-1 possui dezesseis objetos (conjunto de referência de padrões espaciais) que remetem à tipologia de Lambin (Lambin *et al.*, 2003), conforme apresentado na Figura 3-3: geométrico (*GEOM*), corredor (*CORR*), espinha de peixe (*FISH*) e difuso (*DIFF*). Estes objetos foram submetidos ao processo de segmentação e extração de métricas de ecologia da paisagem, cujos dados podem ser verificados na Tabela 4-1. Esta tabela, além das métricas geradas, possui a coluna *ID* que é um identificador de cada objeto processado. A coluna *TIPO* informa a tipologia de cada objeto, a qual é atribuída através de avaliação cognitiva. A partir das métricas e do *TIPO*, o classificador estrutural gerou uma árvore de decisão para cada passo de interação do usuário.

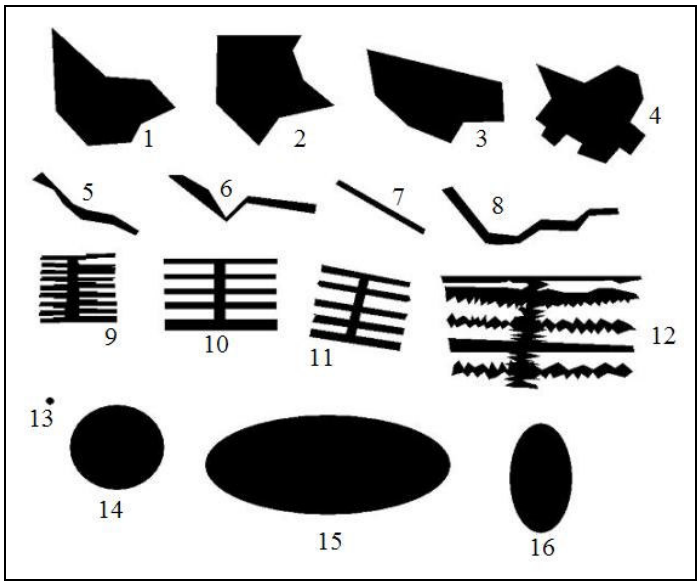


Figura 4-1 - Objetos representando a tipologia de Lambin (Lambin *et al.*, 2003) (da primeira para a última linha): geométrico, corredor, espinha de peixe, difuso

Na primeira interação, o classificador estrutural gerou a árvore de decisão retratada na Figura 4-2. O modelo, utilizando as métricas *SHAPE* e *AREA*, foi capaz de distinguir exatamente cada um dos objetos segundo sua tipologia, ou seja, até determinado limiar *SHAPE* distingue *DIFF* e *GEOM*. Para distinguir *CORR* e FISH, o algoritmo utiliza também *AREA*.

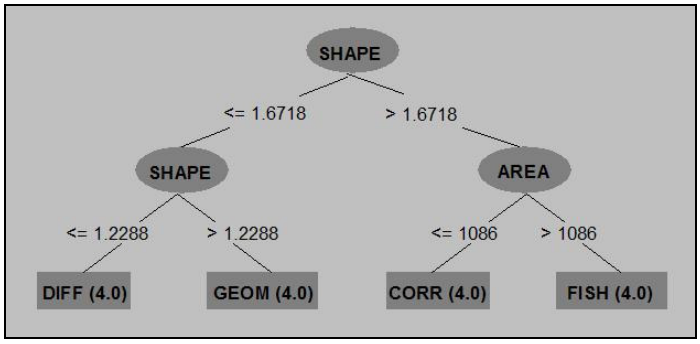


Figura 4-2 – Modelo da primeira interação com o classificador estrutural (*SHAPE/AREA*)

Com o intuito de verificar se o algoritmo seria capaz de distinguir os objetos sem as métricas de *área* e *perímetro*, estas foram excluídas dos dados a serem modelados. Numa segunda interação, o algoritmo gerou a árvore de decisão apresentada na Figura 4-3. Desta vez, apenas a métrica *SHAPE* foi suficiente para distinguir todos os objetos. Para uma melhor verificação deste fato, a Tabela 4-1 apresenta *SHAPE* destacada e em ordem crescente segundo o modelo da Figura 4-3. Observamos então na tabela que cada tipo é identificável, para este conjunto de dados, utilizando-se apenas *SHAPE*.

Tabela 4-1 - Métricas de ecologia da paisagem dos dados sintéticos

ID	AREA	PERIM	GYRATE	PARA	SHAPE	FRAC	CIRCLE	CONTIG	TIPO
14	3657	27400	2275.432	7.4925	1.1322	1.0143	0.0606	0.9894	DIFF
16	3135.25	26600	2258.138	8.4842	1.1875	1.0199	0.4669	0.9877	DIFF
13	27	2500	195.7074	92.5926	1.1905	1.0295	0.2051	0.8735	DIFF
15	11186	52100	4510.776	4.6576	1.2288	1.0225	0.5703	0.9932	DIFF
3	4413.75	35300	2776.082	7.9977	1.3271	1.0323	0.5901	0.9882	GEOM
2	4644.75	36400	2647.481	7.8368	1.3333	1.0328	0.4492	0.9886	GEOM
1	4274.25	37000	2705.46	8.6565	1.4122	1.0395	0.5623	0.9877	GEOM
4	4268	43800	2570.985	10.2624	1.6718	1.0588	0.4648	0.986	GEOM
5	559.5	25800	2108.569	46.1126	2.7158	1.1291	0.911	0.9369	CORR
7	340.5	21900	1900.304	64.3172	2.9595	1.1446	0.9261	0.913	CORR
6	806	34400	3008.035	42.6799	3.0175	1.1394	0.9192	0.9401	CORR
8	1086	40200	3444.703	37.0166	3.0455	1.1377	0.9218	0.947	CORR
10	2504.5	88000	2682.341	35.1368	4.3781	1.1738	0.6914	0.9417	FISH
11	2089.75	85800	2399.062	41.0575	4.6885	1.1834	0.6669	0.9373	FISH
9	2147.75	93200	2038.847	43.3942	5.0108	1.1913	0.5561	0.9289	FISH
12	6181	232000	4119.771	37.5344	7.3651	1.2228	0.699	0.9454	FISH

Numa terceira interação, a métrica *SHAPE* também foi excluída dos dados a serem modelados, gerando o modelo apresentado na Figura 4-4. Este novo modelo foi capaz de classificar corretamente todos os objetos, desta vez utilizando a métrica *FRAC*.

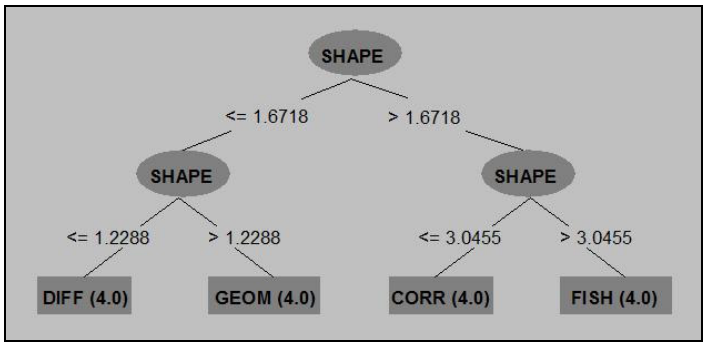


Figura 4-3 – Modelo da segunda interação com o classificador estrutural (*SHAPE)*

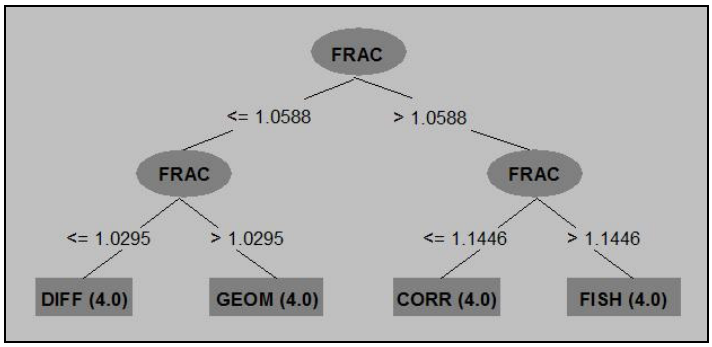


Figura 4-4 - Modelo da terceira interação com o classificador estrutural (*FRAC)*

Outro ponto verificado com os dados sintéticos diz respeito à dependência das métricas em relação a fatores como rotação e extensão dos objetos. Este mesmo conjunto de dados sintéticos teve todos os seus objetos rotacionados; as métricas extraídas deste segundo conjunto possuem, para poucos elementos, diferenças irrelevantes de valores, os quais permanecem inalterados para a maioria dos objetos quando comparados às métricas originais. *SHAPE, FRAC e CIRCLE* são independentes em relação à extensão dos objetos; as demais métricas sofrem variações quando um objeto tem sua extensão alterada, mesmo mantendo-se sua estrutura (forma) original.

Verificamos que a avaliação do usuário nas diferentes fases do processo é fundamental, pois os componentes da metodologia fornecem resultados que podem ser aprimorados através de interações que buscam um melhor ajuste do modelo para a tarefa em questão. Haja vista que os objetos do conjunto de dados sintéticos possuem uma boa definição estrutural, faz-se

necessário verificar o desempenho da metodologia em objetos presentes em imagens de sensoriamento remoto, cujas feições geralmente apresentam variações inerentes à dinâmica da paisagem. Esta avaliação em dados reais é apresentada a seguir.

4.3 Minerando Padrões em Dados de Imagens

Para esta tarefa, foram selecionados 16 objetos da paisagem de imagens do satélite CBERS-2 (INPE, 2005c), sensor CCD, resolução de 20m, cenas 173/111 e 174/111, de junho e agosto de 2004, as quais retratam regiões do estado de Rondônia. Estas imagens foram segmentadas e os objetos selecionados foram cognitivamente associados à tipologia de Lambin, conforme a Figura 4-5. Denominaremos estes dados de *conjunto de referência de padrões espaciais 1* (CRPE 1). Para efeito de apresentação dos dados, suas escalas foram desconsideradas (os objetos dos padrões *espinha de peixe* e *difuso*, por exemplo, possuem grandezas de extensão muito distintas). O processo de extração de métricas de ecologia da paisagem destes objetos resultou nos dados apresentados na Tabela 4-2.

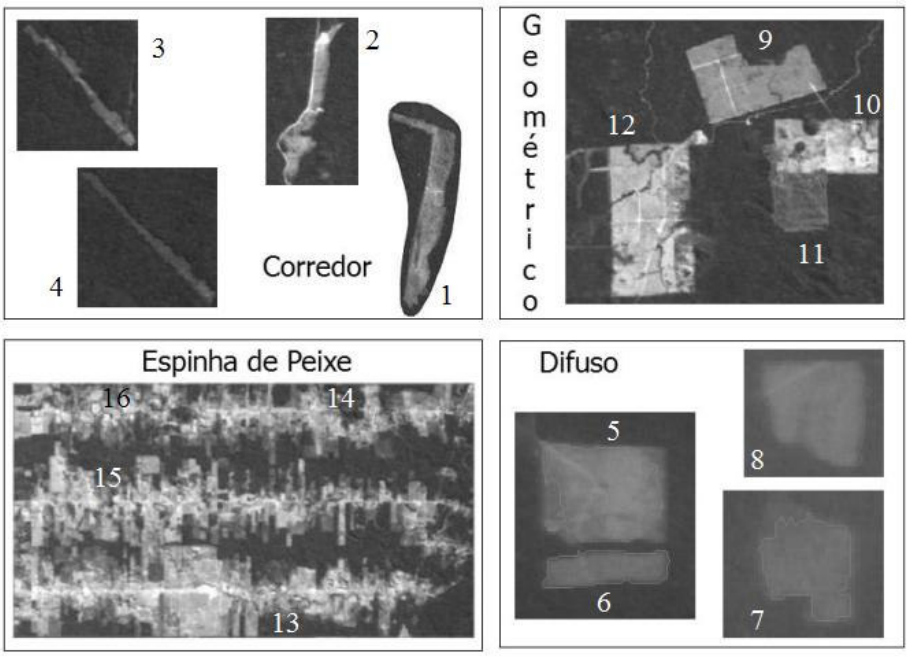


Figura 4-5 - Conjunto de referência de padrões espaciais 1 (CRPE 1)

Interações com o classificador estrutural, a partir das métricas do CRPE 1, geraram o modelo apresentado na Figura 4-6. Este modelo aplicado diretamente ao seu conjunto de dados original consegue classificar os objetos com 100% de acerto utilizando as métricas *AREA* e

SHAPE. Entretanto, uma validação cruzada (cross validation) (Good, 2001) foi aplicada ao modelo para obter um parâmetro estatístico mais rigoroso em relação à árvore de decisão gerada. Nesta validação, 81,25% das amostras foram classificadas corretamente.

Na validação cruzada empregada nos experimentos (denominada *k*-fold, do inglês *k*-partições), o conjunto de dados é particionado em *k* subconjuntos. Num primeiro passo, a primeira partição é isolada. O modelo é treinado utilizando as demais partições e testado com a partição isolada. Num segundo passo, a segunda partição é isolada e o modelo é treinado com as demais partições e testado com a partição isolada, e assim sucessivamente com o objetivo de obter a soma dos erros das *k* iterações e, conseqüentemente, o percentual de amostras classificadas corretamente. O *k* utilizado nos experimentos foi igual a 10, ou seja, todas as validações foram realizadas através de 10 partições.

Tabela 4-2 – Métricas de ecologia da paisagem do CRPE 1

ID	AREA	PERIM	GYRATE	PARA	SHAPE	FRAC	CIRCLE	CONTIG	CLASS
1	386.28	20040	1280.8714	51.8795	2.5431	1.1234	0.823	0.9698	CORR
2	156	13720	951.776	87.9487	2.744	1.1417	0.8786	0.9484	CORR
3	87.8	12840	1036.8538	146.2415	3.4149	1.1799	0.9308	0.9193	CORR
4	40	10080	784.1469	252	3.9375	1.2143	0.9531	0.8592	CORR
5	59.88	4440	299.9656	74.1483	1.4231	1.0542	0.4051	0.9562	DIFF
6	24.52	2760	257.0131	112.5612	1.38	1.0535	0.7157	0.9299	DIFF
7	14.52	2160	149.1492	148.7603	1.3846	1.0587	0.4112	0.9114	DIFF
8	38.16	2920	240.0305	76.5199	1.1774	1.026	0.4262	0.9537	DIFF
9	996.12	21520	1404.4923	21.6038	1.7025	1.0662	0.582	0.9874	GEOM
10	729.76	18400	1149.5012	25.2138	1.6974	1.0674	0.5165	0.9847	GEOM
11	489.52	28200	936.2328	57.6075	3.1757	1.1505	0.7128	0.9658	GEOM
12	1718.08	41680	1781.3448	24.2596	2.5108	1.1107	0.5925	0.9857	GEOM
13	12090.28	379760	5618.7226	31.4104	8.6309	1.2317	0.6893	0.9811	FISH
14	965.68	60440	2367.0821	62.588	4.8585	1.1967	0.8603	0.9629	FISH
15	6258.6	237680	4352.6171	37.9765	7.5025	1.2246	0.776	0.9774	FISH
16	860.36	46840	1341.2394	54.4423	3.983	1.1734	0.6196	0.968	FISH

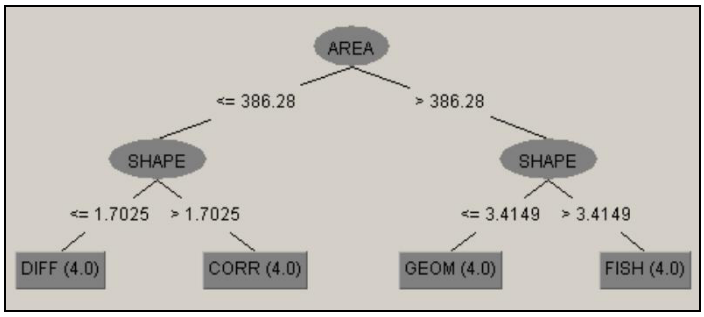


Figura 4-6 – Modelo gerado pelo classificador estrutural para as métricas do CRPE 1

Um segundo conjunto de 16 objetos foi obtido a partir das mesmas imagens CBERS-2/CCD. Denominaremos estes objetos de *conjunto de referência de padrões espaciais 2* (CRPE 2), os quais são retratados na Figura 4-7 (desconsiderar escala). As métricas de ecologia da paisagem extraídas são apresentadas na Tabela 4-3. O modelo gerado pelo classificador estrutural (Figura 4-8) utilizou as métricas *PARA* e *FRAC* para distinguir cada objeto segundo a sua classe com 100% de acerto. A validação cruzada aplicada ao modelo indicou 93,75% de acerto.

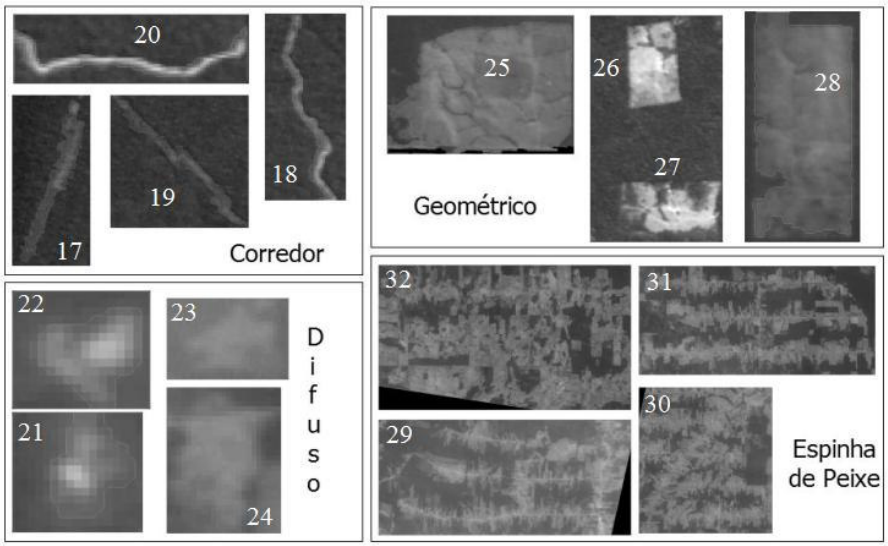


Figura 4-7 - Conjunto de referência de padrões espaciais 2 (CRPE 2)

Tabela 4-3 - Métricas de ecologia da paisagem do CRPE 2

ID	AREA	PERIM	GYRATE	PARA	SHAPE	FRAC	CIRCLE	CONTIG	CLASS
17	9.64	4080	391.5082	423.2365	3.1875	1.2073	0.9429	0.7462	CORR
18	16.52	5960	494.2811	360.7748	3.6341	1.2162	0.9436	0.7914	CORR
19	27.2	8160	567.9395	300	3.8491	1.218	0.9387	0.8324	CORR
20	43.84	7960	562.4789	181.5693	2.9701	1.1694	0.9049	0.8952	CORR
21	3.8800	960.0000	77.6046	247.4227	1.2000	1.0374	0.3601	0.8557	DIFF
22	4.1600	1000.0000	80.6662	240.3846	1.1905	1.0383	0.4292	0.8574	DIFF
23	7.9600	1680.0000	115.1116	211.0553	1.4483	1.0705	0.5385	0.8744	DIFF
24	10.3200	1880.0000	130.9733	182.1705	1.4242	1.0659	0.5204	0.8908	DIFF
25	209.08	8760	632.6604	41.8978	1.5103	1.0571	0.5839	0.974	GEOM
26	145.76	7000	501.1104	48.0241	1.4463	1.0523	0.4852	0.9708	GEOM
27	122.88	10000	501.6583	81.3802	2.2523	1.116	0.6215	0.9516	GEOM
28	814.9200	17680.0000	1111.4224	21.6954	1.5455	1.0549	0.5057	0.9869	GEOM
29	11053.6000	411520.0000	6817.6550	37.2295	9.7795	1.2463	0.8186	0.9782	FISH
30	15212.6800	549080.0000	8322.7344	36.0936	11.1240	1.2558	0.8264	0.9787	FISH
31	18652.0000	452480.0000	6820.3094	24.2591	8.2811	1.2220	0.6409	0.9856	FISH
32	19529.6800	562320.0000	7468.4425	28.7931	10.0558	1.2419	0.7083	0.9831	FISH

Com o intuito de verificar o desempenho das árvores de decisão com um segundo critério de avaliação, cada uma delas foi validada com o conjunto de dados da outra. Desta forma, o modelo gerado para o CRPE 1 (Figura 4-6) foi validado com dados do CRPE 2 (Tabela 4-3),

o que possibilitou classificar corretamente 81,25% das instâncias. Já a árvore modelada para o CRPE 2 (Figura 4-8) foi validada com dados do CRPE 1 (Tabela 4-2), gerando um percentual de 75% de instâncias classificadas corretamente. As matrizes de confusão destas duas validações são apresentadas na Figura 4-9.

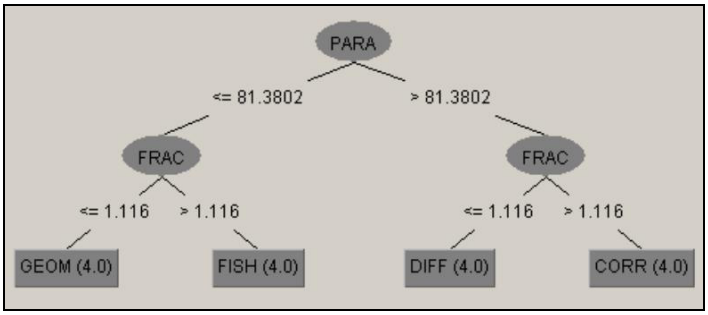


Figura 4-8 - Modelo gerado pelo classificador estrutural para as métricas do CRPE 2

```
== Matriz de confusão: Trein. CRPE 1
                       Valid. CRPE 2

 a  b  c  d   <-- classificado como
 4  0  0  0  | a = CORR
 0  4  0  0  | b = DIFF
 1  2  1  0  | c = GEOM
 0  0  0  4  | d = FISH
```

```
== Matriz de confusão: Trein. CRPE 2
                       Valid. CRPE 1

 a  b  c  d   <-- classificado como
 3  0  0  1 | a = CORR
 0  2  2  0 | b = DIFF
 0  0  3  1 | c = GEOM
 0  0  0  4 | d = FISH
```

Figura 4-9 – Matrizes de confusão da validação mútua entre CRPE 1 e CRPE 2

Os experimentos realizados com dados sintéticos e de imagens através de diferentes interações indicam que: (a) as métricas representam significativamente a estrutura dos objetos; (b) o classificador estrutural é capaz de construir modelos eficazes para distinguir objetos segundo suas métricas; (c) diferentes modelos com bom desempenho de classificação podem ser construídos a partir de subconjuntos distintos de métricas oriundos dos mesmos dados (Figura 4-2, Figura 4-3 e Figura 4-4, por exemplo). Na próxima seção, um modelo construído pelo classificador estrutural será aplicado em imagens de

sensoriamento remoto para a detecção de configurações espaciais, possibilitando assim realizar tarefas de mineração através da metodologia apresentada.

4.4 Mineração de Padrões de Mudança em São Félix do Xingu

Utilizando a metodologia proposta, imagens TM/Landsat 5 (INPE, 2005a) (225/64, 226/64, 226/65, 225/65) de 1997, 2000, 2001, 2002 e 2003, com resolução original de 30 metros e reprocessadas para 60 metros, que cobrem a região de São Félix do Xingu no estado do Pará, foram selecionadas e processadas. As imagens e dados de desflorestamento foram fornecidos pelo Projeto PRODES (INPE, 2005b). A *tipologia de padrões espaciais* para esta tarefa é baseada em conceitos de mudança de uso do solo em florestas tropicais (Tabela 4-4). *Objetos prototípicos* da paisagem extraídos dos dados foram *avaliados e associados cognitivamente* com elementos desta *tipologia de padrões espaciais*, originando um *conjunto de referência de padrões espaciais* com 72 instâncias.

Tabela 4-4 – Mudança de uso do solo em florestas tropicais

Tipologia de padrão espacial	**Mudança de uso do solo**
Linear	Colonização ao longo de estradas Desflorestamento ao longo de rios
Irregular	Agricultura familiar Pequenos incrementos de desflorestamento
Geométrico	Grandes fazendas

O *classificador estrutural*, utilizando o *conjunto de referência de padrões espaciais*, gerou o modelo apresentado na Figura 4-10. O treinamento do modelo foi realizado com 36 instâncias do *conjunto de referência de padrões espaciais*, enquanto o teste do modelo foi efetuado com as demais 36 instâncias. A árvore de decisão obteve 100% de sucesso na classificação dos dados de teste. As métricas que descreveram os padrões espaciais, neste caso, foram *PERIM* e *CIRCLE*. A partir deste modelo, o *classificador estrutural* extraiu *configurações espaciais* das imagens acima citadas, sendo estas utilizadas para abordar pontos relacionados à *dinâmica da paisagem*.

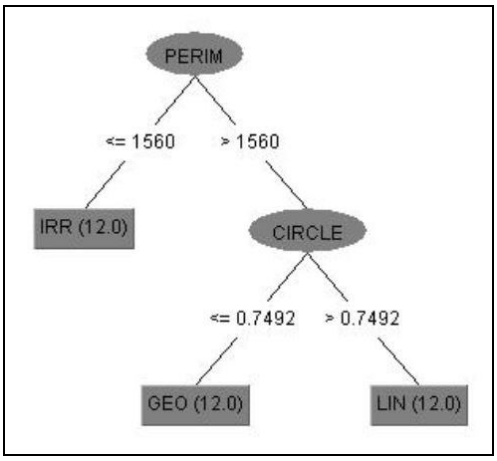


Figura 4-10 – Árvore de decisão - São Félix do Xingu

Num primeiro momento, desejamos responder à seguinte questão: *"Qual o comportamento de grandes fazendeiros em São Félix do Xingu durante o período de 1997 a 2003? A área de novas e grandes fazendas está aumentando?"* Observando a evolução da correspondente configuração espacial para grandes fazendas (*padrão geométrico*) na Figura 4-11, foi possível observar que *"em 2000, este tipo de desflorestamento atingiu um pico de 55.000 ha, mas sofreu redução nos anos seguintes; em 2003, a área desflorestada associada a grandes fazendas diminuiu para 29.000 ha. Tal cenário indica que grandes fazendas estão reduzindo sua contribuição no processo de desflorestamento da região"*.

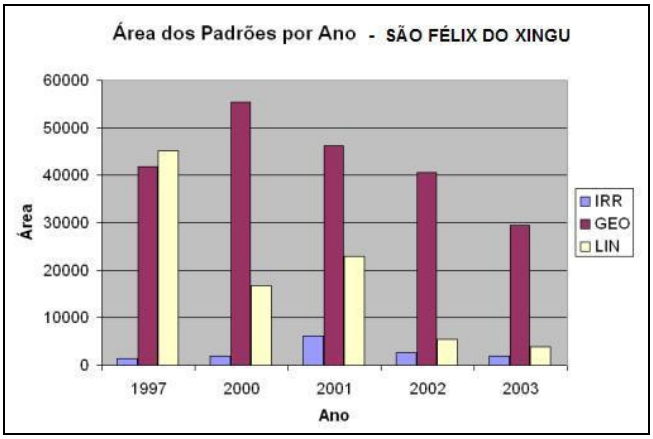


Figura 4-11 – Área (ha) dos padrões espaciais em São Félix do Xingu (1997-2003)

Uma outra questão foi colocada: *"Qual a distribuição dos pequenos agricultores e de pequenos incrementos de desflorestamento nesta área em São Félix do Xingu durante o período de 1997 a 2003?"*. Observando a Figura 4-12, conclui-se que *"a distribuição deste*

padrão de uso do solo neste período concentrou-se principalmente no nordeste e sudeste desta área".

Uma terceira questão seria: *"Nesta região de São Félix do Xingu há algum padrão dominante de mudança de uso do solo?"*. Observando a Figura 4-13, podemos afirmar que *"o padrão irregular representa 61% do total de ocorrências de mudança de uso do solo em 2001, indicando uma intensificação da agricultura familiar e dos pequenos incrementos em áreas desflorestadas neste ano".*

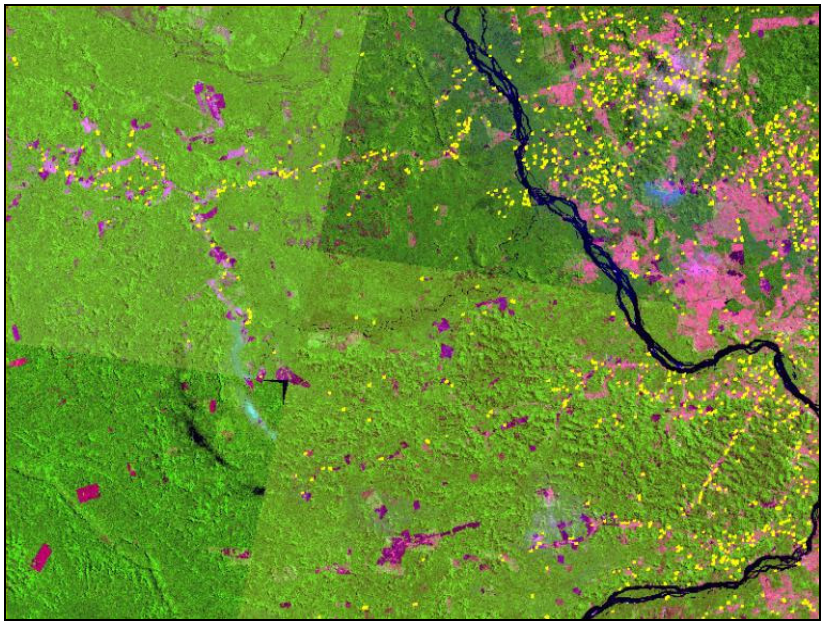


Figura 4-12 - Padrões irregulares (amarelo) em São Félix do Xingu (1997-2003)

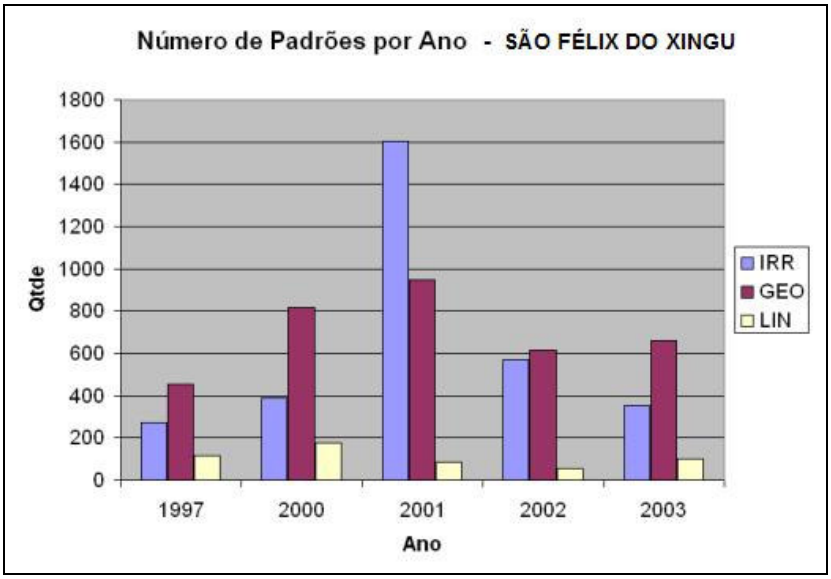


Figura 4-13 – Ocorrência dos padrões espaciais em São Félix do Xingu (1997-2003)

Estes experimentos com dados de imagens em São Félix do Xingu proporcionaram uma abordagem semântica de padrões de mudança através da metodologia proposta. Com o objetivo de obter novas e relevantes informações estratégicas a respeito de processos de desflorestamento em áreas da Amazônia, experimentos foram realizados com dados do Vale do Anari e Terra do Meio, os quais são apresentados nas próximas seções.

4.5 Minerando Padrões de Mudança no Vale do Anari

O "PA Vale do Anari" é um projeto de assentamento planejado pelo INCRA (Instituto Nacional de Colonização e Reforma Agrária), localizado no município de Vale do Anari, estado de Rondônia. Este projeto de assentamento foi estabelecido em 1982, com lotes de terra de aproximadamente 50 ha. Neste caso, o principal processo que desejamos capturar utilizando a mineração de imagens é a *concentração de lotes de terra*. Este processo pode ser descrito como uma aquisição sucessiva de lotes de terra em um assentamento rural planejado pelo governo, resultando em fazendas médias e grandes (Escada, 2003). Processos de concentração de terra são detectados e relatados através de observações em trabalho de campo, os quais são mencionados em vários estudos sobre este processo na Amazônia (Coy, 1987) (Pedlowski; Dale, 1992) (Almeida; Campari, 1995) (Machado, 1998) (Campari, 2002) (Escada, 2003). Para detectar concentração de terra no PA Vale do Anari, Imagens do sensor TM/Landsat 5 (231/66 e 231/67, de 1985 a 2000, com resolução de 30m), tipologia e dados de desflorestamento utilizados nesta tarefa foram gentilmente fornecidos por Isabel Escada.

A tipologia de padrões espaciais de desflorestamento tropical proposta por (Lambin *et al.*, 2003) tenta capturar o processo geral de desflorestamento numa escala mundial. Entretanto, esta tipologia não é adequada para descrever processos de desflorestamento na Amazônia. Em Rondônia, por exemplo, o INCRA desenvolveu diferentes arranjos para organizar lotes de terra para colonos em assentamentos planejados (Escada, 2003). Além disso, para associar padrões espaciais a processos de mudança de uso e cobertura do solo em uma paisagem tão heterogênea e extensa como a Amazônia, é preciso considerar fatores regionais do contexto, como histórico de ocupação e atividades econômicas.

Com o objetivo de detectar a dinâmica dos padrões espaciais de desflorestamento e associá-los a diferentes atores e processos, reconhecemos três estruturas elementares na análise dos dados de desflorestamento (Figura 4-14): irregular, linear e geométrico. A maioria dos

padrões espaciais de desflorestamento na Amazônia pode ser derivada a partir destas estruturas ou da combinação delas. Por exemplo: os padrões espaciais apresentados na Figura 3-3 como *difuso* e *espinha de peixe* são compostos, respectivamente, por desmatamentos irregulares e lineares em diferentes arranjos espaciais, especialmente agregação e ortogonalidade. Características e aspectos semânticos associados aos padrões elementares (irregular, linear e geométrico) no Vale do Anari são apresentados na Tabela 4-5.

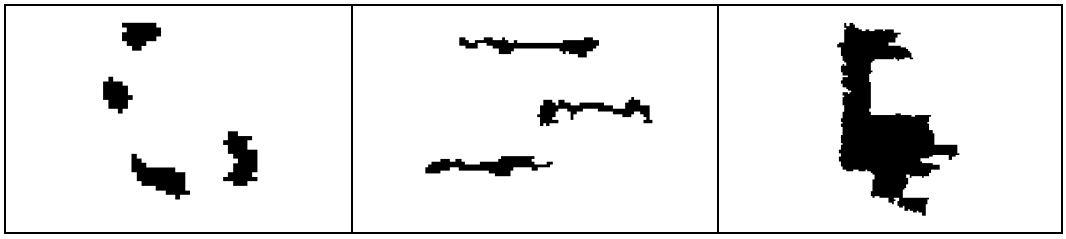


Figura 4-14 – Padrões elementares de desflorestamento (da esquerda para a direita): irregular, linear e geométrico

Tabela 4-5 - Características dos padrões de desflorestamento - Vale do Anari

Tipologia de padrão de desflorestamento	**Distribuição espacial**	**Extensão do desmatamento**	**Atores**	**Principal uso do solo**	**Descrição**
1. Linear	Ao longo de estradas	Variável	Pequenos colonos	Mão-de-obra familiar, agricultura de subsistência e/ou criação de gado	Lote de assentamento do INCRA com 50 ha. Desmatamento ao longo de estradas, com padrão linear seguindo aquelas planejadas pelo INCRA em períodos iniciais da colonização.
2. Irregular	Próximo a estradas principais e núcleos populacionais	Pequeno (< 50 ha)	Pequenos colonos	Mão-de-obra familiar, agricultura de subsistência e/ou criação de gado	Lote de assentamento do INCRA com 50 ha. Desmatamento irregular próximo a estradas seguindo a configuração de lote do INCRA.
3. Geométrico	Próximo a estradas e núcleos populacionais	Médio e grande (> 50 ha)	Médios e grandes fazendeiros	Principalmente criação de gado	Localizado próximo a estradas, seguindo a configuração de lotes do INCRA. Padrões originados da *concentração de lotes*.

Após a extração de *objetos prototípicos* da paisagem, estes foram avaliados e associados *cognitivamente* com *tipos de padrões de desflorestamento* da Tabela 4-5, dando origem a um *conjunto de referência de padrões espaciais* com 46 elementos. A partir deste conjunto, o *classificador estrutural* gerou a árvore de decisão apresentada na Figura 4-15. O treinamento do modelo foi efetuado com todas as instâncias do *conjunto de referência de padrões espaciais*, ao passo que o teste do modelo foi realizado através de validação cruzada (cross-validation), indicando um desempenho de 98% na classificação dos dados, cuja matriz de confusão encontra-se na Figura 4-16. As métricas descritivas dos padrões espaciais derivadas dos passos interativos de mineração foram *CIRCLE* e *PARA*. A partir deste modelo, o *classificador estrutural* extraiu *configurações espaciais* dos dados, permitindo detectar e analisar *concentração de lotes de terra* na *dinâmica da paisagem* do Vale do Anari.

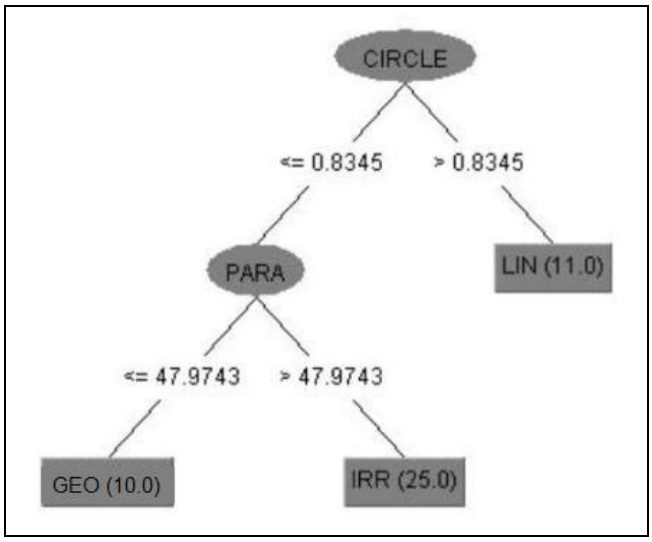


Figura 4-15 - Árvore de decisão – Vale do Anari

```
== Matriz de confusão ==

  a   b   c  <-- classificado como
 10   0   0  |  a = GEO
  0  11   0  |  b = LIN
  0   1  24  |  c = IRR
```

Figura 4-16 - Matriz de confusão da validação cruzada do modelo para o Vale do Anari

Neste estudo de caso, desejamos responder questões relacionadas à evolução de padrões espaciais num típico assentamento do INCRA com processo de concentração de lotes de terra: "*Qual foi o padrão de desmatamento predominante? Quando o processo de concentração de lotes surgiu? Como este evoluiu? Em que proporção este processo ocorre?*".

No início da implantação do assentamento, o padrão de desmatamento dominante foi o *linear*, que corresponde a aberturas ao longo de estradas realizadas por colonos (Figura 4-17). Entretanto, o padrão *irregular* iniciou sua predominância a partir de 1988 seguindo uma lógica de desflorestamento, na qual o ponto de início da abertura é geralmente a frente dos lotes conectados a uma estrada, evoluindo até o limite anterior do lote. Tanto o padrão *linear* como o *irregular* são causados pelos mesmos atores: pequenos colonos com estratégias de uso da terra semelhantes. Em 1988, o padrão *geométrico* pode ser observado, conforme apresentado no gráfico (Figura 4-17) e no mapa de desflorestamento (Figura 4-18). Este padrão iniciou um crescimento progressivo a partir deste ano até 2000. Em 2000, o *geométrico* correspondia a praticamente 30% da taxa de desflorestamento trienal, indicando que lotes do INCRA foram adquiridos de forma ilícita, promovendo concentração de terras.

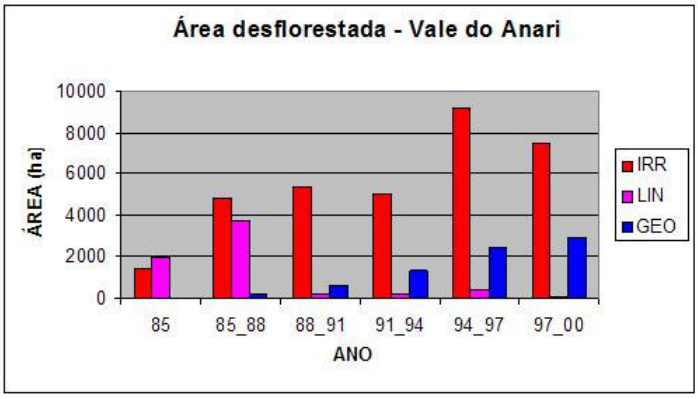


Figura 4-17 – Área dos padrões espaciais no Vale do Anari (1985-2000)

O processo de concentração foi observado durante trabalho de campo efetuado em 2001 (Escada, 2003), os quais foram revelados através de entrevista com a população local e funcionários do INCRA, conforme indicado na Figura 4-19. Os resultados alcançados pela mineração de imagens apresentaram consistência em relação aos dados do trabalho de campo nesta área, demonstrando a potencialidade da metodologia na análise de transformações da paisagem em regiões amazônicas.

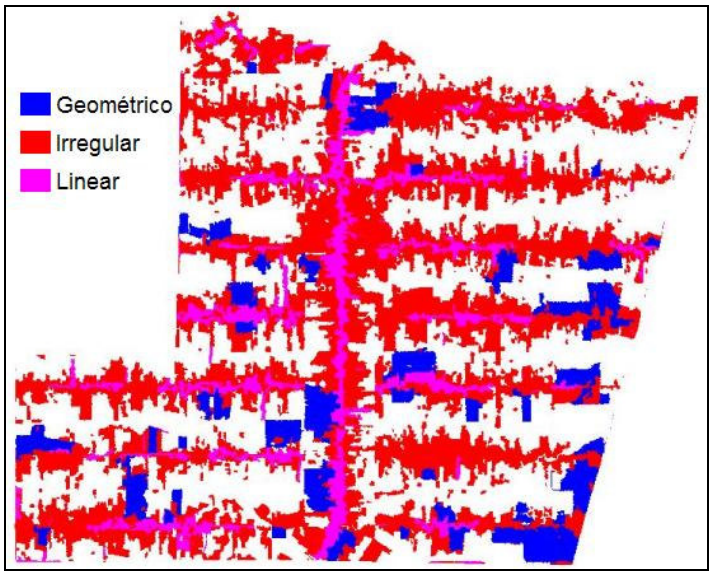


Figura 4-18 – Mapa de padrões de desflorestamento no Vale do Anari (1985-2000)

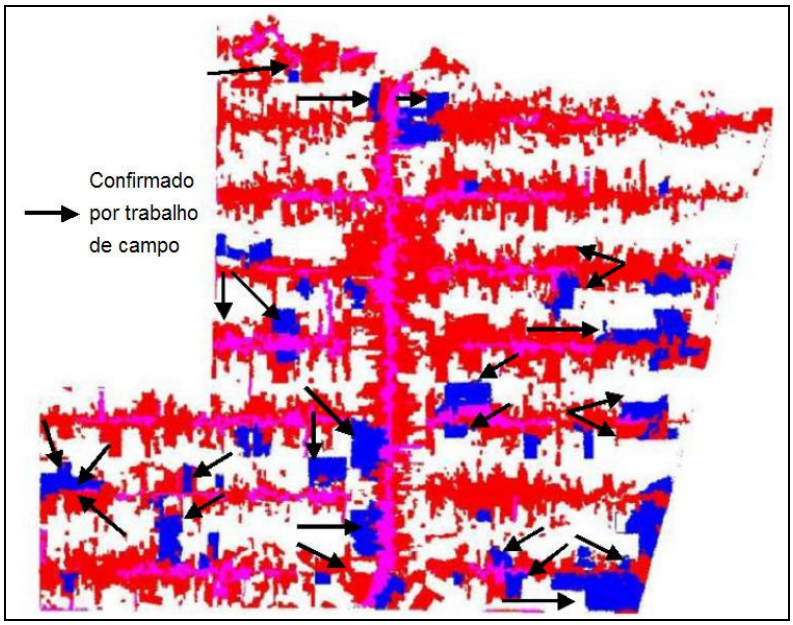


Figura 4-19 – Concentrações de terra confirmadas por trabalho de campo de (Escada, 2003)

4.6 Minerando Padrões de Mudança na Terra do Meio

A região da Terra do Meio (estado do Pará) é uma nova e extensa fronteira de desflorestamento (Becker, 2004), caracterizada por processos ilegais de apropriação de terra e elevado número de grandes fazendas, cujo processo de ocupação não planejado iniciou-se em 2001 (Escada *et al.*, 2005). A tipologia que estabelecemos (Figura 4-20) objetiva detectar atores rurais e sua distribuição espacial ao longo da área de estudo. Esta análise pode fornecer uma melhor compreensão dos papéis dos diferentes atores na transformação da paisagem, bem como da organização espacial dos tipos de fazendas encontrados na região, uma vez que a ocupação da área não foi planejada.

Utilizando a metodologia proposta, dados de desflorestamento fornecidos pelo Projeto PRODES (INPE, 2005b) (de 1997 a 2004, resolução de 60m) foram processados. *Objetos prototípicos* da paisagem extraídos dos dados foram *avaliados cognitivamente* e associados (por um especialista) a *tipos de padrões de desflorestamento* da Tabela 4-6, a qual retrata o desflorestamento como um indicador dos padrões de uso da terra e dos seus atores. O processo deu origem a um *conjunto de referência de padrões espaciais* com 85 instâncias.

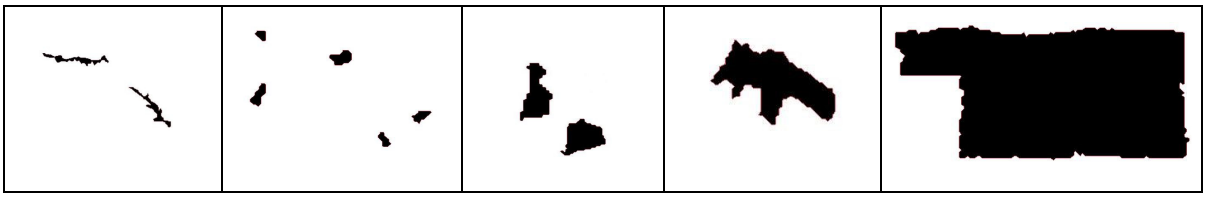


Figura 4-20 – Padrões espaciais de desflorestamento - Terra do Meio (esquerda para direita): linear, irregular pequeno, irregular, geométrico médio, geométrico grande

Através do *conjunto de referência de padrões espaciais*, o *classificador estrutural* gerou o modelo da Figura 4-21. O treinamento do modelo foi realizado com todas as instâncias do *conjunto de referência de padrões espaciais*, enquanto que para o teste foi utilizada validação cruzada, a qual atribuiu 94% de desempenho. Sua matriz de confusão está retratada na Figura 4-22. As métricas determinantes dos padrões espaciais foram *GYRATE, AREA, SHAPE* e *PARA*.

Tabela 4-6 - Características dos padrões de desflorestamento – Terra do Meio

Tipologia de padrão de desflorestamento	Distribuição espacial	Extensão do desmatamento	Atores	Principal uso do solo	Descrição
1. Linear (LIN)	Ao longo de estradas	Variável	Pequenos colonos	Mão-de-obra familiar, agricultura de subsistência e/ou criação de gado	Abertura ao longo de estradas, com o padrão linear acompanhando estradas principais relacionadas ao período inicial de colonização.
2. Irregular pequeno (IPEQ)	Próximo a estradas principais e núcleos populacionais	Pequeno (< 35 ha)	Pequenos fazendeiros e/ou colonos	Mão-de-obra familiar, agricultura de subsistência e/ou criação de gado	Localizado próximo a estradas principais, até a distância de 10 Km.
3. Irregular (IRR)	Próximo a estradas e núcleos populacionais	Pequeno (35 – 190 ha)	Pequenos fazendeiros	Principalmente criação de gado	Localizado próximo a estradas associadas a pequenos colonos. Estes atores geralmente possuem outra fonte de renda, atividades comerciais etc. Utilizam mão-de-obra familiar e de terceiros.
4. Geométrico médio (GMED)	Isolado ou próximo a estradas secundárias	190 – 900 ha	Médios fazendeiros	Criação de gado	Localizado próximo a estradas secundárias associadas a grandes fazendas.
5. Geométrico grande (GGDE)	Isolado ou no término de estradas secundárias	Grande (> 900 ha)	Grandes fazendeiros	Criação de gado	Localizado em regiões isoladas, às vezes próximas a rios. Algumas fazendas possuem pista de pouso.

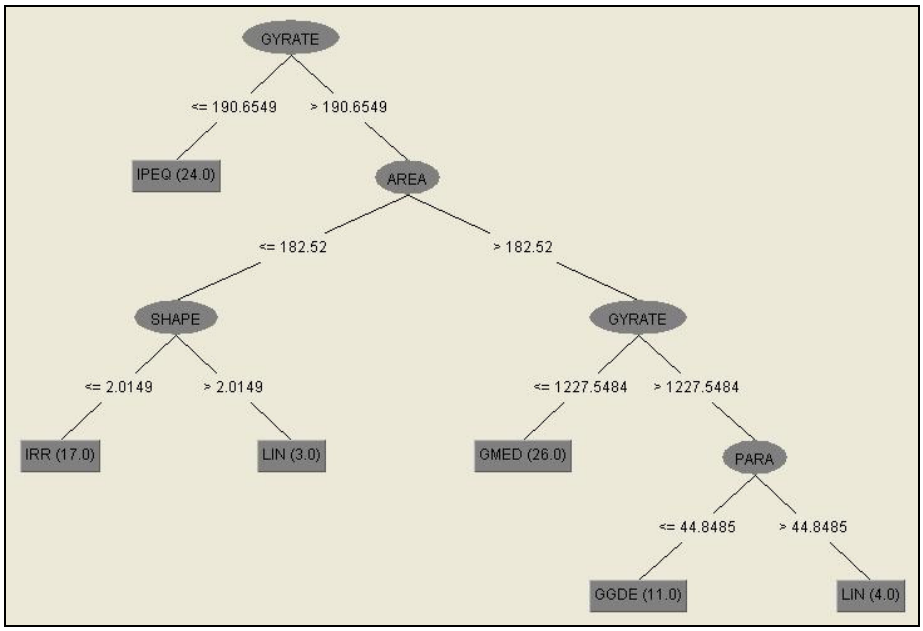


Figura 4-21 - Árvore de decisão – Terra do Meio

```
== Matriz de Confusão ==

  a   b   c   d   e   <--classificado como
 10   1   0   0   0 |  a = GGDE
  1   5   0   1   0 |  b = LIN
  1   0  25   0   0 |  c = GMED
  0   0   1  16   0 |  d = IRR
  0   0   0   0  24 |  e = IPEQ
```

Figura 4-22 - Matriz de confusão da validação cruzada do modelo para a Terra do Meio

As *configurações espaciais* extraídas pelo *classificador estrutural* a partir do modelo permitem responder questões estratégicas: "*Qual o comportamento dos diferentes tipos de fazendeiros durante este período (1997 a 2004)? A área de novas grandes fazendas está aumentando? Como as fazendas estão espacialmente organizadas na região?* ".

Observando a evolução dos padrões de desmatamento na Figura 4-23, foi possível concluir que a taxa de desflorestamento começou a crescer a partir de 2000, alcançando um pico de 40.000 ha no período 2001-2002. A maior contribuição ao desflorestamento no período de 2001 a 2004 veio dos padrões de desmatamento *geométricos* (médios e grandes), conforme apresentado pelo gráfico da Figura 4-23 e pelo mapa da Figura 4-24. Estes padrões estão associados a fazendas grandes e médias, e após 2001 predominaram na região. Podemos observar os primeiros passos da ocupação humana nesta região. Em 1997 o padrão de desmatamento *linear* predominou, sendo fortemente associado à construção de estradas e desmatamento de fazenda ao longo destas. Em 2001 estradas secundárias, ortogonais à estrada principal (Canopus), começaram a expandir-se (Escada *et al.*, 2005). Padrões de desmatamento grandes e médios, representando fazendas grandes e médias, começaram a surgir de forma acelerada próximo a estas estradas e em locais remotos, até 2004. Estes processos originaram uma configuração espacial específica, onde pequenas fazendas de famílias de colonos concentraram-se ao longo de estradas principais, conforme o detalhe da Figura 4-24 apresenta, enquanto grandes e médias fazendas acomodavam-se próximas a estradas secundárias e em locais remotos, menos acessíveis. Novas e grandes fazendas apresentaram crescimento em 2003-2004 com relação ao período anterior, embora a área desmatada em 2003-2004 seja inferior em relação a 2001-2002.

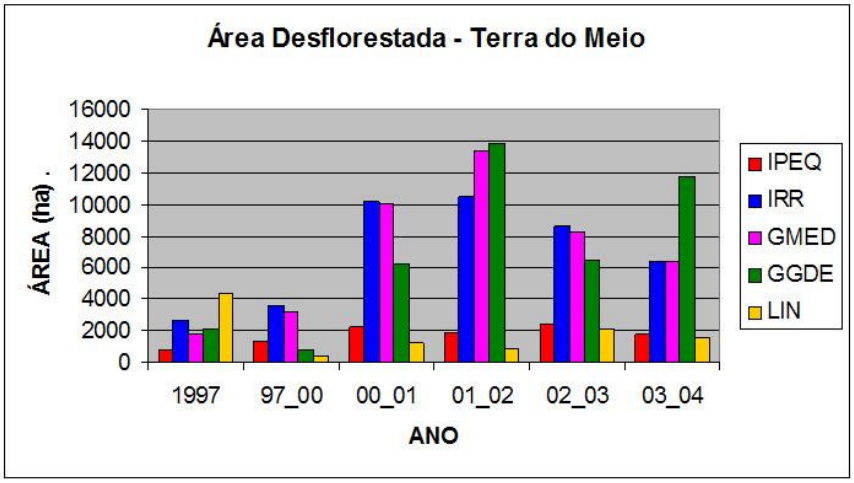


Figura 4-23 - Área dos padrões espaciais na Terra do Meio (1997-2004)

Estes resultados podem ser comparados com observações de trabalho de campo realizadas em 2004, as quais são apresentadas na Figura 4-25. Analisando ambas as figuras e a tipologia de padrões espaciais especificada, verificamos que os padrões de desflorestamento retratados na Figura 4-24 estão relacionados a atores cujo arranjo espacial está apresentado no trabalho de campo da Figura 4-25. Diante da convergência apresentada pelo mapa de padrões gerado em relação ao trabalho de campo (padrões, seus atores e processos subjacentes), podemos constatar que a metodologia é capaz de qualificar padrões de mudança de uso do solo capturando a dinâmica do processo, o que valida os resultados da mineração de imagens realizada.

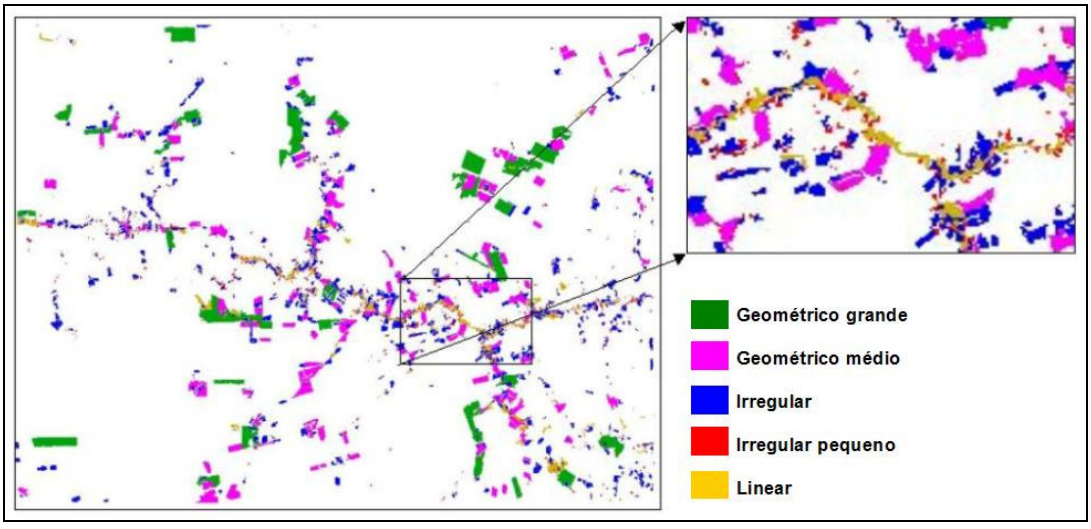


Figura 4-24 - Mapa de padrões de desflorestamento na Terra do Meio (1997-2004)

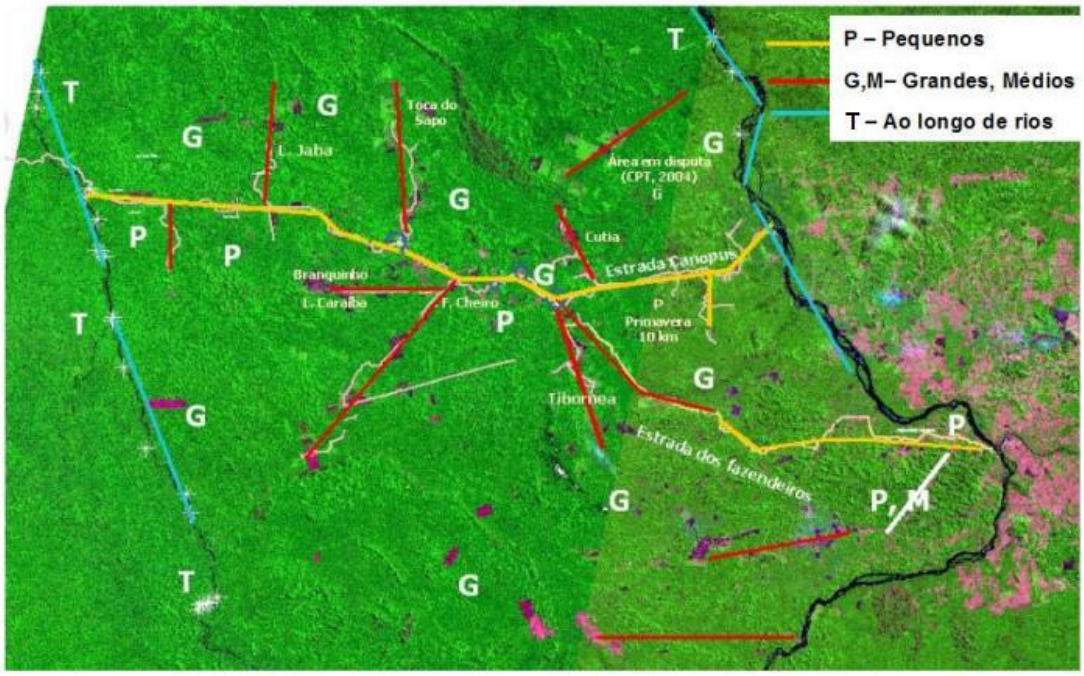


Figura 4-25 – Atores do desflorestamento e sua distribuição espacial na Terra do Meio
FONTE: (Escada *et al.*, 2005)

4.7 Considerações sobre Mineração de Imagens de Sensoriamento Remoto

Um dos aspectos importantes a serem considerados na análise de padrões de desflorestamento, e na tarefa de relacioná-los a processos de mudança de uso do solo, é o fato de que os processos são dinâmicos e evoluem ao longo do tempo (Forman, 1995). Se padrões de desflorestamento forem analisados a partir de dados de uma única data, a observação do processo é gravemente comprometida, uma vez que os padrões observados tendem a ser o resultado da combinação de processos de diferentes períodos, atores e/ou estratégias de uso do solo. Devido a estes fatores, os experimentos realizados utilizaram imagens que retratam a paisagem de diferentes regiões amazônicas em vários períodos, fazendo uso de tipologias e descrições semânticas que caracterizam apropriadamente atores, processos e estratégias, sem perder de vista elementos históricos, sociais e econômicos relevantes.

Uma limitação relevante na classificação estrutural diz respeito à quantidade e qualidade dos objetos prototípicos utilizados para geração do modelo, pois se o número de amostras ou o poder descritivo destas para distinguir diferentes padrões não for adequado, a árvore de decisão gerada classificará incorretamente muitos objetos. Por outro lado, quando no

conjunto de dados de teste estiverem presentes poucas amostras incorretas, o poder de generalização do classificador permitirá mesmo assim obter um bom modelo, indicando a robustez do algoritmo.

Em relação à metodologia em si, foi observada a importância das interações com o especialista do domínio no sentido de ajustar o modelo e identificar problemas como objetos prototípicos inadequados. Já no que diz respeito à validação do modelo, o método a ser utilizado (validação cruzada ou conjunto de teste) depende fundamentalmente do número disponível de bons objetos prototípicos, uma vez que para se trabalhar com dois conjuntos de dados (um para treinamento e outro para teste) faz-se necessário um número suficiente de amostras para popular os dois conjuntos. Um terceiro ponto está relacionado à potencialidade da metodologia para detectar, quantificar e analisar padrões de mudança de uso do solo, haja vista a coerência dos resultados apresentados com trabalhos de campo realizados por pesquisadores da área.

Diante dos resultados apresentados pelos experimentos realizados, foi possível avaliar a metodologia como um todo, testar os componentes do seu protótipo e comprovar a sua eficácia em dados e contextos reais, verificando a sua eficiência na identificação de padrões de mudança e na compreensão dos atores e processos envolvidos. A aplicação da metodologia pode aumentar consideravelmente as chances de detectar, avaliar e reduzir a alta taxa de desflorestamento amazônico, especialmente diante das condições reunidas pelo INPE, o qual agrega dados, conhecimentos e recursos tecnológicos apropriados ao enfrentamento do dilema da Amazônia.

CAPÍTULO 5

CONCLUSÕES

Esta tese traz como contribuição central a metodologia de mineração de imagens de sensoriamento remoto, a qual permite detectar padrões de mudança nesta categoria de dados. A metodologia proposta utiliza técnicas e algoritmos de domínios distintos para obter configurações espaciais que revelem padrões de mudança de uso do solo em imagens de sensoriamento remoto. Da área de processamento de imagens, a metodologia usa a segmentação e classificação por regiões para identificar objetos da paisagem. Já a ecologia da paisagem fornece métricas de regiões, que provêm uma caracterização espacial das áreas desmatadas. No que diz respeito à aprendizagem de máquina, é utilizado um algoritmo de classificação por árvore de decisão, o qual permite construir um classificador estrutural de imagens.

O desenvolvimento e aplicação da metodologia revelaram que alguns fatores permitiram obter bons resultados durante o processo de mineração de imagens, dentre eles: (a) segmentação por crescimento de regiões identifica satisfatoriamente objetos em áreas de floresta retratadas por imagens de sensoriamento remoto; (b) métricas de regiões caracterizam eficientemente objetos da paisagem presentes na imagem; (c) árvores de decisão abstraem e generalizam com bom desempenho e robustez padrões espaciais representados por suas respectivas métricas.

Um protótipo que integra e implementa funcionalidades requeridas pela metodologia foi desenvolvido, através do qual foram mineradas imagens de sensoriamento remoto de regiões da floresta amazônica. Diante da heterogeneidade do contexto amazônico, outra questão relevante (e até certo ponto esperada) é o fato de que o treinamento e aplicação do modelo devem ser realizados em regiões espacialmente semelhantes, ou seja, treinar o classificador estrutural numa determinada região e aplicá-lo em outra com características espaciais diferentes gera resultados inconsistentes.

A abordagem metodológica propõe-se a preencher a lacuna entre imagens de sensoriamento remoto e aplicações que demandam semântica dos dados. Nesta ótica, com o intuito de

analisar e validar a metodologia desenvolvida e o protótipo implementado, experimentos com dados artificiais e reais foram realizados em estudos de casos de desflorestamento na Amazônia em diferentes períodos.

A hipótese de trabalho da tese foi confirmada, pois com base nos conceitos de processamento digital de imagens, mineração de dados e ecologia da paisagem foi desenvolvida uma metodologia de mineração de imagens de sensoriamento remoto. Os resultados obtidos através dos experimentos realizados permitiram avaliar a metodologia proposta, bem como os componentes do seu protótipo de software. As análises comparativas com trabalhos de campo validaram a eficácia da metodologia em dados e contextos reais, indicando sua eficiência na identificação de padrões de mudança de uso do solo e na compreensão dos atores e processos envolvidos.

Um dos futuros trabalhos relacionados a esta tese diz respeito a novos estudos de caso envolvendo regiões amazônicas com diferentes dinâmicas de desflorestamento, uma vez que a imensa área da floresta, a pluralidade de atores e a diversidade de processos promovem diferentes mudanças no uso do solo. Um segundo ponto seria a extensão da metodologia para a mineração de eventos e processos de outros domínios como dinâmica populacional e vigilância territorial, levando em consideração que a metodologia é genérica o suficiente para ser utilizada em outras aplicações que utilizam imagens. Um terceiro ponto é a integração do protótipo desenvolvido ao Terralib (Câmara *et al.*, 2001b), uma biblioteca de classes e funções para desenvolvimento de aplicações em SIG (Sistema de Informação Geográfica), a qual está disponível na internet com código fonte aberto.

Uma das limitações da metodologia diz respeito à quantidade e qualidade dos objetos prototípicos utilizados para geração do modelo na classificação estrutural, pois se o número de elementos e seu poder descritivo para diferenciar padrões não forem adequados, o modelo gerado (árvore de decisão) classificará erroneamente vários objetos. A metodologia também requer uma tipologia de padrões espaciais apropriada, a qual deve caracterizar adequadamente os padrões espaciais e os aspectos semânticos que devem ser detectados no processo. Um outro ponto diz respeito à intensa dinâmica espaço-temporal da região amazônica, o que demanda a avaliação, por parte de um especialista, dos processos que ocorrem na área de estudo durante as diferentes fases interativas da metodologia,

especialmente a seleção dos objetos prototípicos e a interpretação das configurações espaciais.

Durante este trabalho, obtivemos evidências experimentais de que as descrições qualitativas dos padrões não são independentes da extensão das regiões analisadas, pois aspectos como área e perímetro são relevantes para identificar os padrões de mudança de uso do solo. Por exemplo: o padrão *irregular* está geralmente associado à agricultura de subsistência e atividades minoritárias, cujas áreas são "pequenas"; já o padrão geométrico está relacionado a desflorestamentos em larga escala para agroindústrias e outros empreendimentos, cujas áreas são tipicamente "grandes". Durante os experimentos observamos também quais categorias de métricas são necessárias para diferenciar padrões estruturais: enquanto nos dados sintéticos métricas que descrevem a forma (*SHAPE* e *FRAC*) foram suficientes para caracterizar os padrões, nos dados de imagens as métricas que retratam dimensão das áreas (especialmente *AREA* e *PERIM*) também foram essenciais à criação dos modelos dos padrões. Isto se deve em grande parte à dependência de métricas de extensão que os padrões de mudança de uso do solo possuem, bem como à irregularidade das feições retratadas nas imagens.

O processo de mineração de imagens mostrou-se interativo por natureza, haja vista a necessidade de seleção de amostras, construção e avaliação de modelos, avaliação do cenário, retorno a pontos específicos do processo, dentre outros. Durante os experimentos, a avaliação de resultados em fases distintas indicou a necessidade de novos objetos prototípicos, de melhor calibração do modelo ou mesmo ajustes na tipologia de padrões espaciais.

A expectativa é que a utilização dos recursos deste trabalho aumente consideravelmente o poder de detecção, avaliação e redução do acelerado processo de desflorestamento amazônico, haja vista o *know-how* e a cobertura espaço-temporal de imagens da floresta que o INPE possui.

REFERÊNCIAS BIBLIOGRÁFICAS

Aksoy, S.; Koperski, K.; Tusk, C.; Marchisio, G. Interactive Training of Advanced Classifiers for Mining Remote Sensing Image Archives. In: ACM International Conference on Knowledge Discovery and Data Mining, 2004, Seattle, WA. ACM, p. 773-782.

Almeida, A. l. O. d.; Campari, J. S. **Sustainable Settlement in the Brazilian Amazon**. New York: The World Bank, 1995. 185 p.

Alves, D. Space-time Dynamics of Deforestation in Brazilian Amazônia. **International Journal of Remote Sensing**, v. 23, n. 14

Alves, D.; Escada, M. I. S.; Pereira, J. L. G.; Linhares, C. d. A. Land use intensification and abandonment in Rondônia, Brazilian Amazônia. **International Journal of Remote Sensing**, v. 24, n. 4, p. 899-903

Becker, B. **Amazônia**. São Paulo: Editora Ática, 1997. 112 p.

___. **Amazônia - Geopolítica na Virada do III Milênio (Amazonia - Geopolitics on the Verge of the Third Millenium).** Rio de Janeiro: Garamond, 2004. 167 p.

Bins, L.; Fonseca, L.; Erthal, G. Satellite Imagery Segmentation: a region growing approach. In: VIII Brazilian Symposium on Remote Sensing, 1996, São José dos Campos, BR. INPE, p. 677-680.

Briassoulis, H., 2000, **Analysis of Land Use Change: Theoretical and Modeling Approaches**, Regional Research Institute, WVU.

Câmara, G.; Egenhofer, M.; Fonseca, F.; Monteiro, A. M. What´s In An Image? In: Montello, D. (Ed.). **Spatial Information Theory: Foundations of Geographic Information Science. International Conference, COSIT 2001.**, v. Lecture Notes on Computer Science 2205. Santa Barbara, CA.: Springer, 2001a, p. 474-487.

Câmara, G.; Souza, R.; Freitas, U.; Garrido, J. SPRING: Integrating Remote Sensing and GIS with Object-Oriented Data Modelling. **Computers and Graphics**, v. 15, n. 6, p. 13-22 Disponível em: www.dpi.inpe.br/gilberto.

Câmara, G.; Vinhas, L.; Souza, R. C. M.; Paiva, J. A.; Monteiro, A. M. V.; Marcelo Tílio de Carvalho, B. R. Design Patterns in GIS Development: The Terralib Experience. In: III Simpósio Brasileiro de Geoinformática, 2001b, Rio de Janeiro.

Campari, J. S. **Challenging the Turnover Hypothesis of Amazon Deforestation: Evidence from Colonization Projects in Brazil**. 330 p. Tese de Doutorado em Filosofia (The University of Texas, Austin, 2002.

Chen, Y.; Wang, J. Z.; Krovetz, R. CLUE: Cluster-based Retrieval of Images by Unsupervised Learning. In: Seventh International Symposium on Signal Processing and its Applications, 2003, Paris. IEEE,

Clifton, C. Change Detection in Overhead Imagery using Neural Networks. **International Journal of Applied Intelligence**, v. 18, p. 215--234

Coy, M. Rondônia: Frente pioneira e programa POLONOROESTE: O processo de diferenciação sócio-econômica na perifieria e os limites do planejamento público. **Tubinguen Geographhische Studien**, n. 95, p. 253-270

Escada, M. I. S. **Evolução de Padrões da Terra na Região Centro-Norte de Rondônia**. 164 p. Tese de Doutorado em Sensoriamento Remoto (Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2003. (INPE-10209-TDI/899).

Escada, M. I. S.; Vieira, I. C. G.; Amaral, S.; Araújo, R.; Veiga, J. B. d.; Aguiar, A. P. D.; Veiga, I.; Oliveira, M.; Pereira, J. L. G.; Filho, A. C.; Fearnside, P. M.; Venturieri, A.; Carriello, F.; Thales, M.; Carneiro, T. S.; Monteiro, A. M. V.; Câmara, G. Padrões e Processos de Ocupação nas Novas Fronteiras da Amazônia: O Interflúvio do Xingu/Iriri (Land Use Patterns and Processes in Amazonian New Frontiers: the Xingu/Iriri Region). **Estudos Avançados**, v. 19, n. 54, p. 9-23

Fayyad, U.; Piatesky-Shapiro, G.; Smyth, P.; Uthurusamy, R. **Advances in Knowlege Discovery and Data Mining**. Cambridge: MIT Press, 1996. 560 p.

Fonseca, F.; Egenhofer, M.; Agouris, P.; Câmara, G. Using Ontologies for Integrated Geographic Information Systems. **Transactions in GIS**, v. 6, n. 3, p. 231-257

Forman, R. T. T. **Land Mosaics: The Ecology of Landscapes and Regions**. Cambridge: Cambridge University Press, 1995. 652 p.

Geist, H. J.; Lambin, E. F. **What Drives Tropical Deforestation?** Louvain-la-Neuve: 2001. 116 p. (4).

Good, P. I. **Resampling Methods: A Practical Guide to Data Analysis** Birkhauser Verlag AG 2001. 256 p.

INPE. **Banco de Imagens Digitais Georeferenciadas do Sensor MSS** São José dos Campos, 2005a. Disponível em: www.dgi.inpe.br/catalogomss. Acesso em: 28/20/2005.

___. **Projeto PRODES - Monitoramento da Floresta Amazônica Brasileira por Satélite**. São José dos Campos, 2005b. Disponível em: www.obt.inpe.br/prodes. Acesso em: 19/09/2005.

___. **Satélite Sino-Brasileiro de Recursos Terrestres - CBERS**. São José dos Campos, 2005c. Disponível em: www.cbers.inpe.br. Acesso em: 14/10/2005.

Kohavi, R.; Quinlan, R. Decision Tree Discovery. In: Klosgen, W.; Zytkow, J. M. (Ed.). **Handbook of Data Mining and Knowledge Discovery**. Oxford University Press, 2002, p. 282-288.

Lambin, E. F.; Geist, H. J.; Lepers, E. Dynamics of land-use and land-cover change in Tropical Regions. **Annual Review of Environment and Resources**, v. 28, p. 205-241

MacDonald, J. The Earth Observation Business and the Forces that Impact It. In: Earth Observation Business Network 2002, 2002, Vancouver, CA. MacDonald Dettwiler,

Machado, L. A Fronteira Agrícola na Amazônia. In: Becker, B. K.; Christofoletti, A.; Davidoch, F. R.; Geiger, R. P. P. (Ed.). **Geografia e Meio Ambiente no Brasil**. 1998, p. 181-217.

Marr, D. **Vision: A Computational Investigation into the Human Representation and Processing of Visual Information**. New York: Henry Holt & Company, 1982.

McGarigal, K. Landscape pattern metrics. In: El-Shaarawi, A. H.; Piegorsch, W. W. (Ed.). **Encyclopedia of Environmentrics**. v. 2. Sussex, England: John Wiley & Sons, 2002, p. 1135-1142.

McGarigal, K.; Marks, B. **FRAGSTATS: spatial pattern analysis program for quantifying landscape structure**. Washington, DC: USDA Forestry Service Technical Report PNW-351, 1995.

Meinel, G.; Neubert, M. A comparison of segmentation programs for high resolution remote sensing data. **International Archives of Photogrammetry and Remote Sensing**, v. XXXV, n. Part B, p. 1097-1105

Nagao, M.; Matsuyama, T. **A Structural Analysis of Complex Aerial Photographs**. New York: Plenum Press, 1980. 199 p.

NASA, N. A. a. S. A. **MODIS - Moderate Resolution Imaging Spectroradiometer**. Washington, 2005. Disponível em: http://modis.gsfc.nasa.gov/about. Acesso em: 19/10/2005.

Pedlowski, M. A.; Dale, V. H. **Land use practices in Ouro Preto d'Oeste, Rondônia, Brazil**. Oak Ridge: Oak Ridge National Laboratory, 1992. 12 p. (ORNL Technical Manuscript 3850).

Quinlan, R. **C4.5: Programs for Machine Learning**. San Francisco: Morgan Kaufmann, 1993.

Rushing, J.; Ramachandran, R.; Nair, U. J.; Graves, S. J.; Welch, R.; Lin, A. ADaM: A Data Mining Toolkit for Scientists and Engineers. **Computers and Geosciences, In Press, Available online 11 January 2005**

Schröder, M.; Rehrauer, H.; Seidel, K.; Datcu, M. Interactive Learning and Probabilistic Retrieval in Remote Sensing Image Archives. **IEEE Trans. on Geoscience and Remote Sensing**, v. 23, n. 9, p. 2288--2298

Shimabukuro, Y.; Batista, G.; Mello, E.; Moreira, J.; Duarte, V. Using shade fraction image segmentation to evaluate deforestation in Landsat Thematic Mapper images of the Amazon region. **International Journal of Remote Sensing**, v. 19, n. 3, p. 535-541

Silva, M. P. S.; Câmara, G.; Escada, M. I. S.; Souza, R. C. M. d. **Remote Sensing Image Mining: Detecting Patterns of Change**. Universidade do Estado do Rio Grande do Norte / Instituto Nacional de Pesquisas Espaciais, 2005a.

Silva, M. P. S.; Câmara, G.; Souza, R. C. M.; Valeriano, D. M.; Escada, M. I. S. Mining Patterns of Change in Remote Sensing Image Databases. In: The Fifth IEEE International Conference on Data Mining, 2005b, Houston.

Turner, M. G. Landscape Ecology: The effect of Pattern on Process. **Annual Review of Ecology and Systematics**, v. 20, p. 171-197

U.S. Department of the Interior, U. S. G. S. **EROS, Earth Resources Observation and Science**. Washington, 2005. Disponível em: http://edc.usgs.gov. Acesso em: 19/10/2006.

Witten, I. H.; Frank, H. **Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations**. San Francisco: Morgan Kaufmann, 1999.

Zhang, J.; Hsu, W.; Lee, M. Image Mining: Trends and Developments. **Journal of Intelligent Information**, n. Special Issue on Multimedia Data Mining

Zucker, S. W. Region growing: childhood and adolescence. **Computer Graphics and Image Processing**, v. 15, p. 382-399

APÊNDICE A

ARTIGO SUBMETIDO A REVISTA INTERNACIONAL: INTERNATIONAL JOURNAL OF REMOTE SENSING

Remote Sensing Image Mining: Detecting Patterns of Land Use and Land Cover Change in Tropical Forest Areas

MARCELINO PEREIRA DOS SANTOS SILVA[1,2], GILBERTO CÂMARA[2*], MARIA ISABEL SOBRAL ESCADA[2], RICARDO CARTAXO MODESTO DE SOUZA[2]

[1]UERN – Rio Grande do Norte State University, BR 110, Km 48, 59610-090, Mossoró, RN, Brazil

[2]INPE – National Institute for Space Research, P.O. Box 515, 12201-097, São José dos Campos, SP, Brazil

Abstract

Remote sensing image databases are the fastest growing archives of spatial information. However, we still have a limited capacity for extracting information from large remote sensing image databases. There are few techniques for image data mining and information extraction in large image data sets, and thus we are failing to exploit our large remote sensing data archives. This paper proposes a method for mining land use patterns in remote sensing image databases. The basic idea is to use a structural classifier to describe shapes found in land use maps extracted from remote sensing images, and then associate these shape descriptions to the different types of social actors involved in land use change. We support our proposal with two case studies for detecting land use patterns in Amazonia from INPE's remote sensing image database.

1 INTRODUCTION

Despite the large success of global remote sensing programs and the widespread availability of remotely sensed data, there is a "knowledge gap" when extracting information from images. This "knowledge gap" has arisen because our capacity to build sophisticated earth observation satellites is not matched by our means of producing information from these data sources (MacDonald, 2002). To a significant extent, we are failing to exploit the potential of

the spatial data we collect. One of the areas where this "knowledge gap" is critical is on the use of our large remote sensing image archives. The US National Satellite Land Remote Sensing Data Archive holds 1,400 TB of satellite data collected over a 40-year period, and satellites such as NASA's Terra and Aqua produce an extra three terabytes of imagery daily. Brazil's National Institute for Space Research (INPE) has more than 130 Terabytes of image datasets, covering 30 years of remote sensing activities, which are available on a database with free on-line access for Brazilian researchers.

The availability of large remote sensing image archives leads to a need for techniques for exploiting them. Currently, we have a limited capacity for extracting information from remote sensing image databases. A large remote sensing image database is a collection of snapshots of landscapes, which provide us with a unique opportunity for understanding how, when, and where changes take place in our world. For example, INPE's image database covers a 30-year history of land use change in the Amazon tropical forest. Extensive fieldwork also points out the different actors involved in land use change (small-scale farmers, large plantations, cattle ranchers) can be distinguished by their different spatial patterns of land use (Lambin et al., 2003). Besides, these patterns evolve in time; new small settlements emerge and large farms increase their agricultural area at the expense of the forest. In these and similar cases, patterns of land use change will have similar spectral signatures and knowledge extraction techniques based on clustering in the feature space would not be able to distinguish between them. Therefore, we consider the key problem in information extraction on remote sensing image databases is *tracking patterns of land use change*.

Given this perspective, this paper proposes a method for mining patterns of change in remote sensing image databases. The basic idea is to build generic descriptions of patterns in remote sensing images, and then use structural approaches to identify these patterns in the image database. Our approach builds on by earlier works by our research group on using ontologies for integrated GIS (Fonseca et al., 2002), and on ontological characterization of remote sensing imagery (Câmara et al., 2001). This paper is an extended and revised version of earlier results (Silva et al., 2005). In what follows, we discuss previous work in remote sensing image mining and propose a method for mining land use patterns in remote sensing image databases. We support our proposal with a case study for detecting land use patterns in Amazonia from INPE's remote sensing image database.

2 REMOTE SENSING IMAGE MINING: AN OUTLINE

Given a large remote sensing image database, researchers would like to explore the database with questions such as: What are the different land use patterns present in the database? When did a certain land use pattern emerge? What are the dominant land use patterns for each region? How do patterns emerge and change over time? The answers to these and similar questions require data mining techniques that are able to perform similarity searches between patterns found in different images. We propose to approach this problem by using spatial patterns as a means of describing relevant semantic features of an image.

2.1 General perspective

The proposed method for remote sensing image mining is presented in Figure 1**.** Initially, we select images from a *repository* according to the application needs. The pre-processing phase includes *geometric and radiometric calibration* to improve data quality. Next, the images go through a *feature extraction* procedure, using an object-oriented segmentation algorithm. The results of the segmentation are identifiable regions in the image with defined boundaries. The data mining phase consists of assigning a description to these regions and identifying those that match to a specific pattern of land use change. The results of the data mining phase are analysed according to their spatio-temporal trends. For example, the results may point out an increase of cattle ranching activities during the last five years in a specific area.

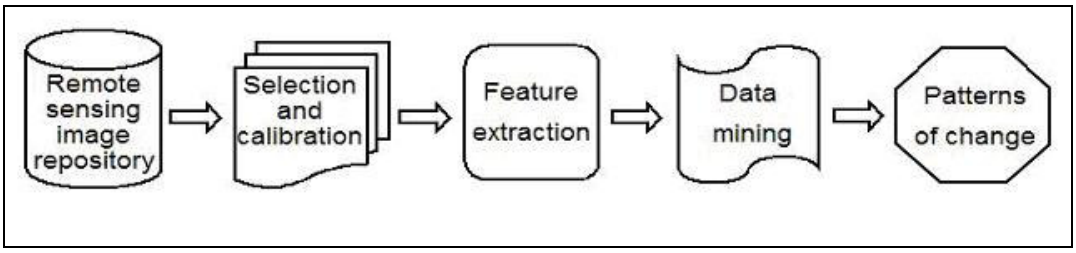


Figure 1. Remote sensing image mining process.

Remote sensing image mining requires an understanding of the differences between remote sensing data and other images. Remotely sensed images are ontologically instruments for capturing landscape dynamics (Câmara et al., 2001). A geographic landscape is an ever-changing scenario, and remote sensing data collection produces images that capture *snapshots* of change trajectories. The challenge for image mining techniques is describing

continuous land use change based on these snapshots. Tracking the temporal evolution of patterns in remote sensing imagery requires methods that are different from standard content-based image retrieval (CBIR) systems. A typical CBIR system uses a query image as the source and images in the database as targets, and query results are a set of images sorted by feature similarities with respect to the source (Chen et al., 2003). When searching for patterns in remote sensing image databases, a different approach is necessary. Instead of similarity searches between image pairs, a system for image mining in remote sensing image databases must be able to describe and extract land use patterns found in different images. Therefore, image mining in remote sensing image databases is searching for patterns of change, not searching for internal content.

2.2 Previous work on remote sensing image mining

Remote sensing image mining systems such as KIM (Datcu et al., 2003), VisiMine (Aksoy et al., 2004) and ADaM (Rushing et al., 2005) focus on methods that work on the feature space. These techniques are useful for distinguishing spectral signatures of different land use types, such as finding areas that are classified as “lakes”, “cities” or “forests”. They are not designed to extract patterns of land use change from multitemporal data.

KIM is a system for processing multisensor image sequences (optical and radar), employing stochastic techniques (Datcu et al., 2003). In KIM, user supplied information is used to build a Bayesian classification network. Features extracted from images on a data set are matched to spectral signatures derived from user choices. The core of the system is a texture-based feature selection procedure based on stochastic models (such as Gibbs-markov random fields). The capabilities of the system are useful for information extraction in SAR images (Datcu et al., 2003). KIM is a sophisticated stochastic mining system, not specifically focused on mining patterns of change in a series of remote sensing images from the same area.

The VisiMine system is a decision-tree classifier (Aksoy et al., 2004). Its rules are based on an entropy-maximization algorithm. The idea is to select, from a set of input data, the features that best distinguish the different objects in the data set. VisiMine offers both pixel-level and region-level classification. In the first case, it allows integrating ancillary data such as digital terrain models with remote sensing images. The region-based classification techniques use basic shape parameters such as eccentricity and rotation. Unlike

our proposal, VisiMine does not use shape parameters that capture the differences between land use patterns.

The ADaM (Algorithm Development and Mining) system is a set of scientific data mining tools (Rushing et al., 2005). It is a collection of general-purpose pattern analysis and image processing techniques, implemented for a grid computing architecture, which supports a wide variety of data formats. The user interface allows building workflows for specific applications. ADAM does not include shape measures that would be suitable for detection of land use patterns shares the advantages and disadvantages of all workflow-based techniques.

3 MINING PATTERNS OF CHANGE IN REMOTE SENSING IMAGES

In this section, we describe our proposed method for extracting patterns of change from remote sensing images. The method considers that instruments onboard remote sensing satellites capture energy at different parts of the electromagnetic spectrum, which is then converted into digital imagery. These instruments are not designed for a specific application, but are a compromise between sensor technology and requirements from different user communities. As a result, remote sensing images have a structural description that is independent of the application domain a scientist employs to extract information. Therefore, we need to distinguish between the image domain and the application domain, as shown in Figure 2:

- *Spatial Patterns* – the geometric structures extracted from the images using techniques for feature extraction, segmentation, and image classification. They are identified and labelled according to a typology that expresses their semantics. Examples of such patterns include *corridor-like regions* and *regular-shaped polygons* representing patterns of the mined data.
- *Application Concepts* – the different classes of spatial objects, associated to a specific user domain. For example, in deforestation assessments, concepts include *large-scale agriculture, small-scale agriculture, cattle ranching* and *wood logging*.

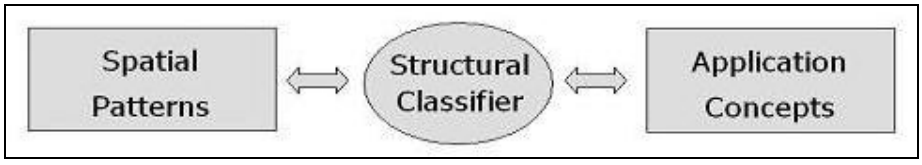


Figure 2. Overview of pattern mining process

To associate structures found in the images to concepts in the application domain, we use a *structural classifier*. Different *structural classifiers* will produce different associations between spatial patterns and the user domain concepts, and that each association is valid within a given application context. Our method consists of three steps:

- Definition of a *spatial pattern typology* according to the user's application domain.
- Building a reference set of *spatial patterns*, using prototypical images.
- Mining the database using a *structural classifier* (guided by the *application concepts* of the domain), matching the reference set of *spatial patterns* to the *objects* identified in the images.

3.1 Defining a spatial pattern typology

The first phase of the method calls for defining a spatial pattern typology which is associated to a given application domain. To illustrate our proposal, we will present typologies defined for mapping different types of land use change in tropical forests.

Remote sensing images are useful for understanding the forces driving land use change in tropical forests. The assumption is the changes in land use can be captured by the spectral and spatial properties of the images (Alves et al., 2003). Extensive fieldwork also points out the different actors involved in land use change (small-scale farmers, large plantations, cattle ranchers) can be distinguished by their different patterns of land use (Mertens and Lambin, 1997). The authors propose a typology of the land use patterns associated to deforestation in tropical forests (Figure 3). Their typology includes *corridor* (commonly associated with riverside and roadside colonization), *diffuse* (related to smallholder subsistence agriculture), *fishbone* (typical of planned settlement schemes), and *geometric* (linked to large-scale clearings).

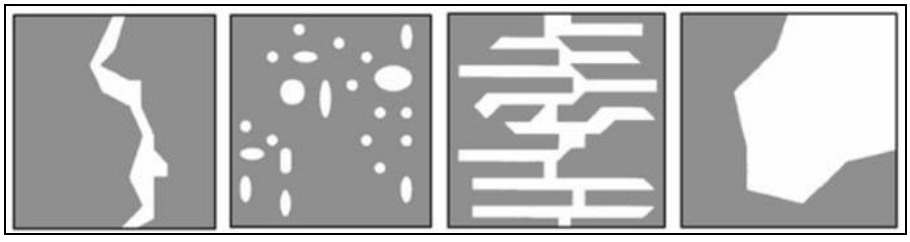


Figure 3. Spatial pattern typology of tropical deforestation (from left to right): corridor, diffuse, fishbone, and geometric (source: Mertens and Lambin (1997)

The spatial patterns typology proposed by Mertens and Lambin (1997) tries to capture tropical deforestation in a worldwide scale. This typology is not satisfactory to describe deforestation process in specific regions such as the Brazilian Amazônia. As an example, in the Brazilian state of Rondônia, there are many fishbone patterns associated with colonization. However, there are other spatial patterns associated to colonization. The Brazilian government used different spatial arrangements to organize colonist land parcels in planned settlement schemes. Colonist land parcels use different spatial arrangements, including fishbone, radial, corridors and dendritic (following geomorphologic features) patterns (Batistella et al., 2003) (Escada et al., 2005). When moving from a global scale to a regional scale, it is better to avoid generalizations such as the proposal by Mertens and Lambin (1997). Associating land use and land cover change patterns with social actors in Amazonia region requires an understanding of occupation history, economic activities, and social and environmental constraints. The analysis of spatial patterns of deforestation also should consider their temporal evolution (Forman, 1995). A specific spatial deforestation pattern results from the land use strategies of different actors. The analysis of spatial deforestation patterns using remote sensing images from a single date does not reveal the whole deforestation process. The spatial patterns detected are the result of a combination of different actions from different actors and their land use strategies. In Section 4, we present case studies in Amazonia where the spatial pattern typology considers the history of each study area.

3.2 Building a reference data set of spatial patterns

In this section, we consider the problem of building a reference set of spatial patterns. For mining land-use patterns in remote sensing images, the application specialist needs to define a typology of spatial patterns that will contain information about the modelled domain. These patterns must match a specific real-world activity and must be distinguishable by an automated procedure. Once the user fixes a typology of spatial patterns for the study area, he needs to associate its concepts with prototypical examples. These prototypes consist of idealized generalizations of land use types, such as the *corridor*, *diffuse*, *fishbone* and *geometric* patterns proposed by Mertens and Lambin (1997). The user needs to find real examples in objects extracted from remote sensing images.

To represent the structures detected in remote sensing images, we introduce the idea of a *landscape object*. A *landscape object* is a structure detected in a remote sensing image by an image segmentation algorithm. *Landscape objects* will be associated to concepts of a spatial pattern typology to characterize them. Figure 4 shows examples of landscape objects that match concepts of a spatial pattern typology, discussed in section 4.

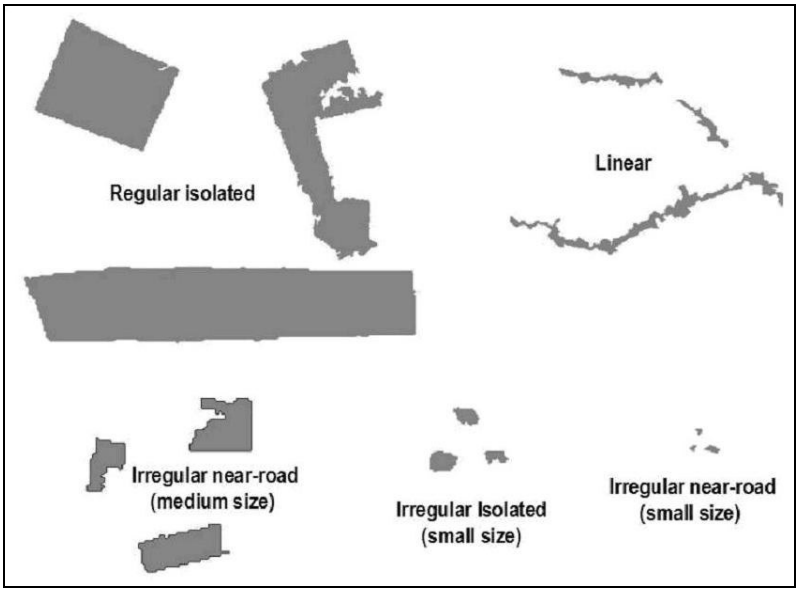


Figure 4. Landscape objects associated to spatial patterns

The steps of building a reference set of spatial patterns are outlined in Figure 5. The first step is to select a set of sample images that capture the land use history of the study area. From this sample, we extract a set of prototypical *landscape objects*. An expert associates these prototypical *landscape objects* to concepts of a *spatial pattern typology,* resulting in a *reference set of spatial patterns* for the study area.

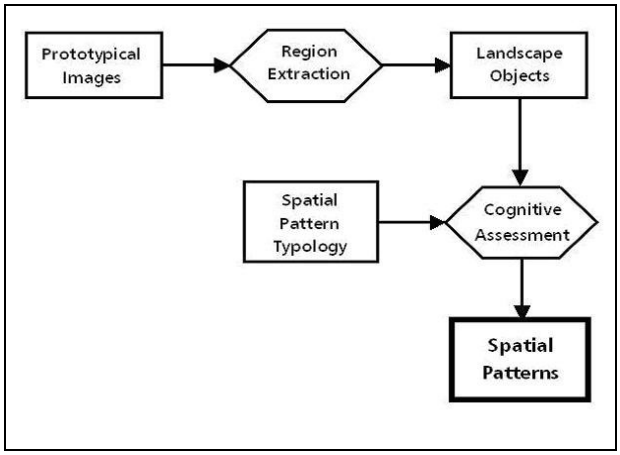


Figure 5. Building a reference set of spatial patterns

To extract landscape objects from remote sensing images, we use segmentation algorithms to partition the image into regions that are spatially continuous, disjoint and homogenous. Recent surveys (Meinel and Neubert, 2004) point out that region-growing approaches (Zucker, 1976) produce closed and homogeneous regions. In our proposal, we have adopted the region-growing segmentation algorithm developed by INPE (Bins et al., 1996), and included in the SPRING software system (Câmara et al., 1996), which is freely available on the Internet. This algorithm has been extensively validated for extracting land use patterns in tropical forests (Shimabukuro et al., 1998) and was favourably reviewed in a recent survey (Meinel and Neubert, 2004). SPRING's region growing algorithm works as follows. Initially, the algorithm breaks the image in segments of one or a few pixels. Then, it compares each segment to its neighbours. Two neighbours merged if they are similar. Each segment continues to grow by comparing it with all the neighbours until there is no remaining joinable region, at which point the algorithm labels the segment as a completed region. The algorithm moves to the next uncompleted cell, repeating the entire sequence until it labels all cells. The algorithm requires two parameters: a similarity threshold and an area threshold.

3.3 Mining the database using a structural classifier

In this section, we describe how to extract information from a remote sensing image database. The method needs two inputs: the reference set of spatial patterns (built as described in the previous section) and land cover maps extracted from the remote sensing images. The land cover maps should be built by an object-oriented remote sensing classification procedure. For each image, the classification procedure extracts landscape objects and labels them according to a set of land cover classes. This procedure is typical of object-oriented image classification (Geneletti and Gorte, 2003). Using the concepts introduced in the previous section, the object-oriented image classification procedure builds a set of *landscape objects*. Using the reference set of *spatial patterns* (see the preceding section), we associate each landscape object to a type of land use pattern. We refer to the set of labelled landscape objects as a *spatial configuration*. To obtain a spatial configuration from a set of landscape object, we use a *structural classifier*. The role of the structural classifier is to assign a spatial pattern to each landscape object. For example, suppose we are using the spatial pattern typology proposed by Mertens and Lambin (1997), pictured in

Figure 3. Then, each landscape object will be associated to either one of a *corridor*, *diffuse*, *fishbone* or *geometric* pattern (see Figure 6).

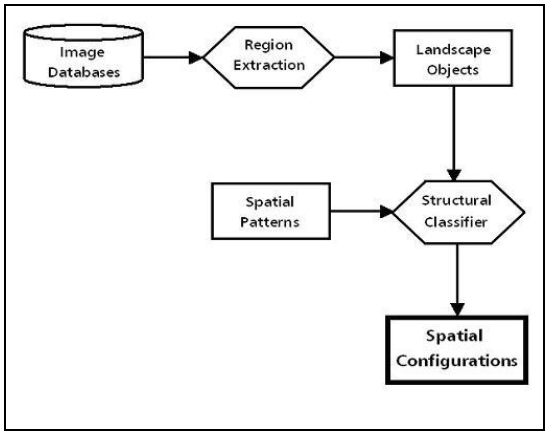


Figure 6. Obtaining spatial configurations

The structural classifier enables the association between landscape objects and concepts in the spatial patterns typology. The structural classifier distinguishes between different spatial patterns. This problem can be mapped into a classification method based on a decision tree that, based on non-categorical attributes, predicts correctly the value of a categorical attribute (Witten and Frank, 1999). The categorical attribute is the pattern type and the non-categorical attributes are a set of numerical values that characterize each pattern. The chosen algorithm was the C4.5 decision tree classifier (Quinlan, 1993). The basic ideas behind the C4.5 classifier are:

- In the decision tree each node matches to a non-categorical attribute and each arc to a possible value of that attribute. A leaf of the tree specifies the expected value of the categorical attribute for the records described by the path from the root to that leaf.
- In the decision tree at each node should be associated the non-categorical attribute which is *most informative* among the attributes not yet considered in the path from the root.
- *Entropy* is used to measure how informative is a node.

To select the attributes that distinguish the different types of land use patterns, we used ideas from *Landscape Ecology* (Turner, 1989). Landscape ecology is based on the notion that environmental patterns strongly influence ecological processes. One of the key items of landscape ecology theory are metrics that characterize geometric and spatial

properties of categorical map patterns (McGarigal, 2002). The pattern metrics used in landscape ecology include metrics of spatial configuration at the patch level. Patches form the building blocks for categorical maps. Patch metrics refer to the spatial character and arrangement, position, or orientation of patches within the landscape. We have used the pattern metrics proposed by the FRAGSTATS software (Spatial Pattern Analysis Program for Categorical Maps) (McGarigal and Marks, 1995), that include:

- *Perimeter (m) and area (ha).*
- *Perimeter-area ratio* (*para*): a measure of shape complexity.
- *Shape* (shape index): patch perimeter divided by the minimum perimeter possible for a maximally compact patch of the matching patch area.
- *Fractal dimension index*: two times the logarithm of patch perimeter (m) divided by the logarithm of patch area (m^2).
- *Circle* (related circumscribing circle): 1 minus patch area (m^2) divided by the area (m^2) of the smallest circumscribing circle.
- *Contiguity index*: equals the average contiguity value for the cells in a patch.

The landscape ecology metrics of the reference set of spatial patterns (as in Figure 3) are fed into the C4.5 classification algorithm. The algorithm builds a decision tree that uses these metrics to distinguish the different types of patterns. After this classifier has been properly trained, it labels the landscape objects found in land use maps. For each land cover map on the image database, this procedure builds a set of labelled landscape objects (called a *spatial configuration*). By identifying the *spatial arrangements* on different images, the user will be able to evaluate the emergence and evolution of different types of change. Each spatial pattern is associated to a different type of land use change. Therefore, the comparison between spatial configurations of images in different locations and between spatial configurations of images at the same location in different times will allow new insights into the actors that bring about change.

4 CASE STUDIES: IMAGE MINING FOR DEFORESTATION PATTERNS

This section presents two case studies of remote sensing image mining for detecting patterns of land use change in tropical forest. Our case studies focus on the analysis deforestation on

the Amazon tropical forest, which covers about 40% of the Brazilian territory. The causes of deforestation include economic, social and political ones. The current pace of land use change is substantial, with an average of 25,000 km^2 of forest being cleared every year. We have used the method described in Section 3 to obtain a better understanding of the processes of land use change in Amazonia, assessing the role and the spatial organization of the different actors involved in land use change. We present two case studies that show the use of data mining image techniques in Amazônia. The first study case is the *Terra do Meio* region *(São Félix do Xingu* and *Altamira* municipalities) in the *Pará* state (Figure 7). There, deforestation has increased in the last 5 years, associated to unplanned occupation. The second study case is a planned rural settlement in the Vale do Anari municipality in state of Rondônia (Figure 8).

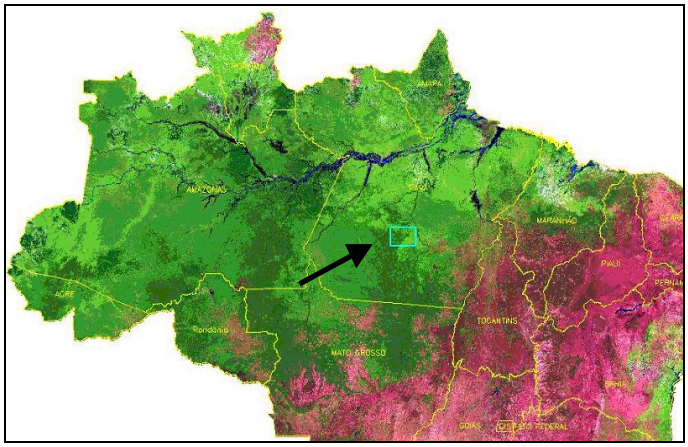


Figure 7. Terra do Meio case study area

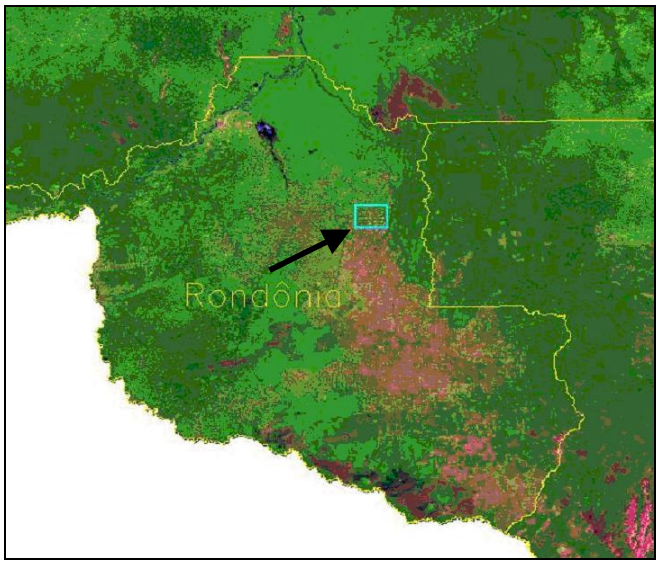


Figure 8. Vale do Anari case study area

4.1 The "Terra do Meio" case study

The "Terra do Meio" region is a large area in the state of Pará (Becker, 2004), where much public land has been seized by illegal procedures. The deforestation rate increased strongly in the period of 2000 to 2004. The area has large farms (many established by illegal means) and small settlers associated with migration. There are five types of social actors in the area:

- Small households associated with migrant families, who live on subsistence agriculture or work for the farmers. Their land use pattern is associated to roadside colonization and shows up as linear patterns in the land cover maps.
- Small farmers and family households that live out near the main roads or close to population settlements. Their land use patterns show up in the maps as small, irregular patches.
- Small cattle ranchers that live near to roads or to settlements. These ranchers are distinguished by (35-190 ha) irregular land use patterns of size.
- Farmers with medium-sized properties (190-900 ha) that are isolated or close to secondary roads, and show up in the maps as medium-sized regular patterns.
- Farmers with large-sized properties (more than 900 ha) that are usually isolated and located close to rivers, and that show up in the maps as large-sized regular patterns.

Table 1 presents the typology for land use actors and Figure 9 shows examples of the five spatial patterns (*linear, small irregular, irregular, medium regular and large regular*).

The prototypical landscape objects were extracted from deforestation maps (INPE, 2005) for the period from 1997 to 2004, with 60m of spatial resolution. Some examples of these prototypical objects are shown in Figure 9. The 1997-2000 period has three years of temporal resolution, while the 2000-2004 period has one year of temporal resolution. We obtained a reference set of spatial patterns based on the deforestation clearings as indicators of land use patterns and actors. The *structural classifier*, using the *spatial patterns*, extracted *spatial configurations* from the set of maps just mentioned. We wanted to answer the following question: *"What's the behaviour of different types of farmers from 1997 to 2004? Is the area of new large farms increasing? How the farms are spatially organized in the region?"* Based on the prototypical objects, the structural classifier associated each landscape object found in the maps to one of the land cover classes described in Table 1. The

distribution of types of clearing patterns is presented in Figure 10 and shows the evolution of human occupation in this region.

Table 1. Linking Clearing Patterns with Land Use Change Process

Land use patterns	Spatial distribution	Clearing size	Actors	Main land use	Description
1. Linear	Roadside	Variable	Small household	Family labor, subsistence crop and/or cattle ranching	Roadside clearings, with linear pattern following main roads corresponding to the earlier stages of colonization.
2. Small Irregular	Near main roads and populational nucleus	Small (< 35 ha)	Small farmers and/or family household	Family labor, subsistence crops and/or cattle ranching	Located near main roads (Canopus and Fazendeiros Road), up to the distance of 10 Km
3. Irregular	Near roads and populational nucleus	Small (35-190 ha)	Small farmers	Cattle ranching mainly	Located near roads, associated to small family household. These actors often have another incoming source from salary, commercial activities, etc. They use family and external labor.
4. Medium Regular	Isolated or near secondary roads	190-900 ha	Medium farmers	Cattle ranching	Located near secondary roads, associated to large farms.
5. Large Regular	Isolated or at the end of secondary roads	Large (> 900 ha)	Large farmers	Cattle ranching	Located in isolated region, sometimes near rivers. Most of them has airstrip.

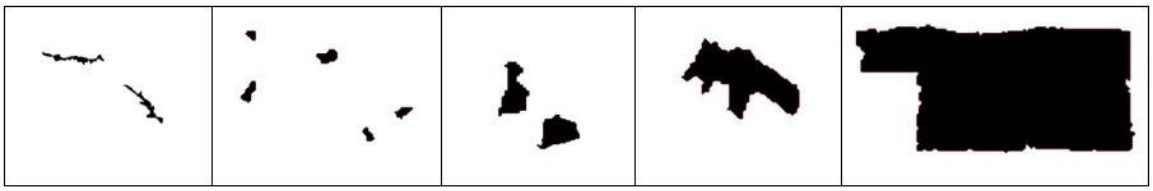


Figure 9. Examples of spatial patterns typology in Terra do Meio: (from left to right) linear, small irregular, irregular, medium regular, large regular.

In Terra do Meio, the deforestation rate started to increase after 2001 and reached a peak of 40,000 ha in the period of 2001 to 2002. In 1997, the linear clearing pattern predominated, associated to road construction and roadside farm clearings. The most important contribution to deforestation rates from 2001 to 2004 came from large and medium geometric clearing patterns, associated with large and medium farms. As the detail

of Figure 11 shows, the resulting spatial configuration has small farms and family households concentrated along main roads, and large and medium farms arranged near secondary roads and in remote places.

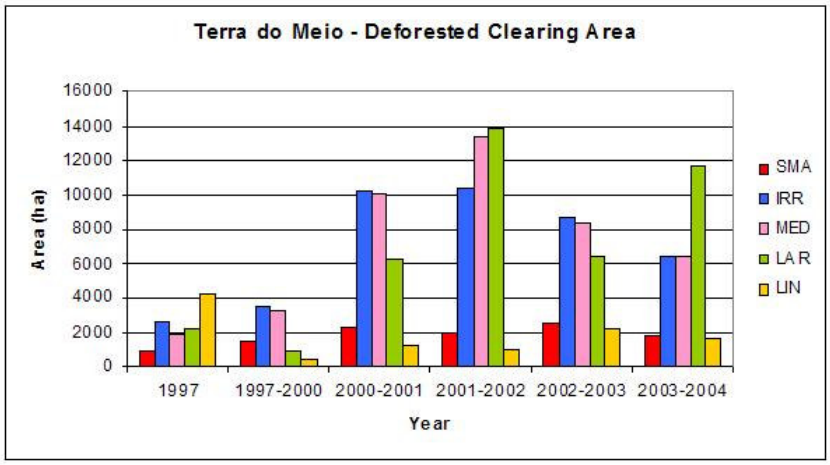


Figure 10. Distribution of clearing patterns in Terra do Meio (1997-2004)

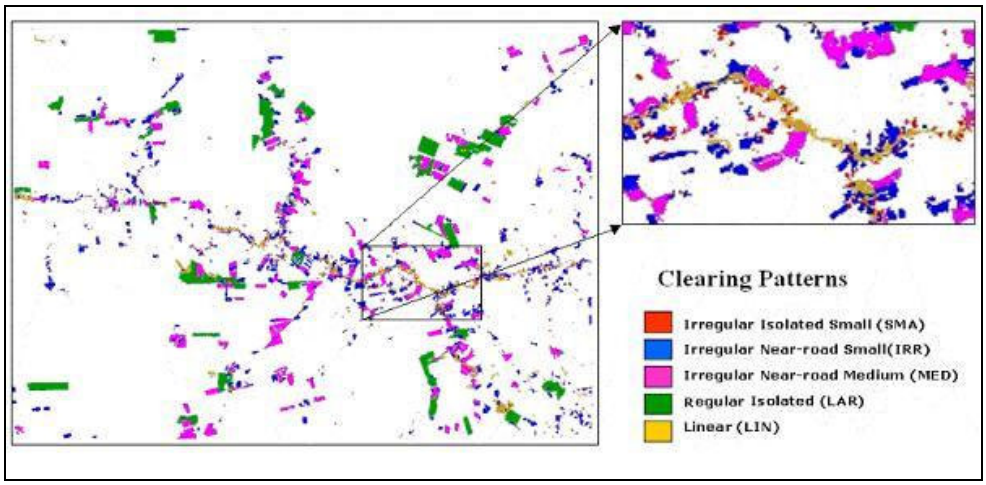


Figure 11. Clearing patterns in Terra do Meio (1997-2004)

We confirmed these results with fieldwork carried out in 2004 (Escada et al, 2005). The spatial configuration obtained using image data mining techniques agrees strongly with fieldwork observations. This lead us to conclude that image data mining is a powerful tool to quantify and to analyze land cover change in new frontier areas, where data is scarce and the pace of change is fast. Drivers behind the landscape transformation can be detected and associated with spatial and temporal land use cover change patterns.

4.2 Rondônia Case Study

The second case study was a small-scale planned rural settlement in Vale do Anari municipality in the state of Rondônia. This settlement was established in 1982 and land parcels sized around 50 ha. We wanted to capture the process of land parcel concentration using spatial data mining. Land concentration results from merging of many land parcels in a rural settlement, where one farmer buys the parcels from the original settlers. This results in farms with medium to large size. The land concentration process in Rondônia and other regions of rural settlement in Amazônia has been reported and detected by fieldwork observations (Almeida and Campari, 1995) (Campari, 2002) (Escada et al., 2005). To detect land concentration, clearing attributes such as size and shape must be considered. We considered three types of social actors in the Vale do Anari area, associated with three different spatial patterns:

- Pre-settlement household colonists living on subsistence agriculture or small cattle ranching. Their spatial patterns show up as *linear* patterns following planned roads corresponding to the earlier stages of colonization.
- Small household colonists associated to settlement schemes living on subsistence agriculture or small cattle ranching. Their spatial patterns show up as *irregular* clearings near roads, following parcels defined by the planned settlement.
- Medium to large farmers, associated to cattle ranches larger than 50 ha. Their spatial patterns are *regular* ones, close to roads and population nucleus.

Table 2 presents the land use and actor typology associated to spatial patterns. In this study, we wanted to answer questions related to the land concentration process on a typical rural settlement, including: *"What was the predominant clearing pattern in the settlement? How did it evolve? When did land concentration in the settlement started to emerge? In which proportion has this process happened?"* The prototypical landscape objects were extracted from deforestation maps (Escada et al., 2005) for the period from 1985 to 2000, with 30m of spatial resolution and three year intervals. Figure 12 shows the clearing patterns in Vale do Anari from 1985 to 2000 and Figure 13 the distribution of spatial pattern types for the same period.

Table 2. Spatial patterns and the associated typology to Vale do Anari

Land use patterns	Spatial distribution	Clearing size	Actors	Main land use	Description
1. Linear	Roadside	Variable	Small household colonist	Family labor, subsistence crop and/or cattle ranching	Settlement scheme – 50 ha land parcel. Roadsided clearings, with linear pattern following planned roads corresponding to the earlier stages of colonization.
2.Irregular	Near main roads and population nucleus	Small (< 50 ha)	Small household colonist	Family labor, subsistence crops and/or cattle ranching	Settlement scheme - 50 ha land parcels. Irregular clearing near roads following parcels configuration.
3. Regular	Near roads and population nucleus	Medium and large (> 50 ha)	Medium to large farmers	Cattle ranching, mainly	Located near roads, following parcels configuration. Regular pattern originated from concentration of more than one parcel.

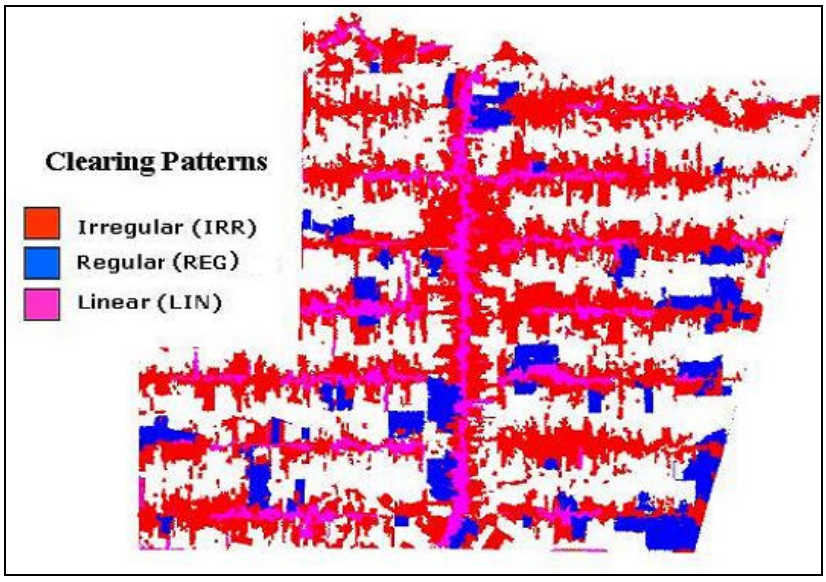


Figure 12. Clearing patterns in Vale do Anari (1985-2000)

The distribution shown in Figure 13 signals a concentration of land ownership. In the earlier stages of the rural settlement, the dominant clearing patterns were linear and irregular. The linear patterns match the roadside clearings of household colonists and the irregular patterns also result from deforestation caused by colonists who cleared their land parcels.

Both clearing patterns correspond to the land use strategies of colonists in different occupation stages. From 1998 onwards, regular patterns emerge and grow progressively, as shown in the pattern distribution (Figure 13). The regular patterns increase in time to reach almost 30% of the deforestation in the period 1987-2000. This shows a marked land concentration process, showing the government plan for settling many colonists in the area has been partially frustrated. Large farmers have bought the parcels in an illicit way, promoting land concentration. Land concentration was confirmed by fieldwork in the region (Escada et al., 2005), showing the method supports the analysis of the landscape transformation in different scales, applications and regions of Amazônia.

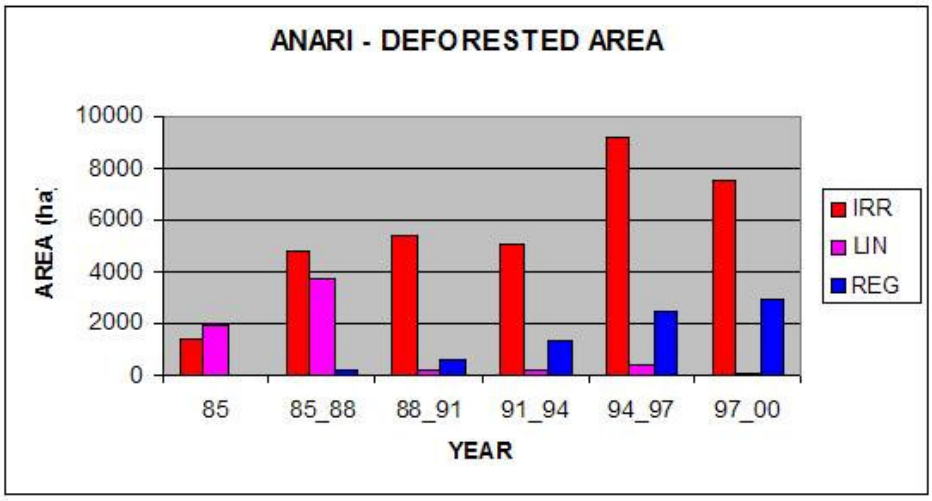


Figure 13. Distribution of clearing patterns in Vale do Anari (1985-2000)

5 CONCLUSIONS

This paper proposes a method for mining patterns of change that enables extracting spatial arrangements from remote sensing image databases. This method addresses the problem of describing land use change. It combines techniques from data mining, digital image processing and landscape ecology to describe shapes on the maps resulting from remote sensing image classification. Structural pattern classification in maps extracted from images of distinct dates enables associating land change objects to causative actors. The results from the case studies show that pattern classification techniques associated to remote sensing image interpretation are a step forward in understanding and modelling land use change. The proposed method also enables a more effective use of the large remote sensing image databases available in agencies such as USGS, ESA and INPE. Further experiments are

necessary to improve the method, to test alternatives for image segmentation algorithms and for structural classifiers. Experimental evidence shows that qualitative description of land use patterns is scale dependent: attributes such as area and perimeter are relevant to identify them. Future research directions in remote sensing image mining include tracking individual trajectories of change. Patterns found in one map would be linked to those in earlier and later maps, thus enabling an explicit description of the trajectory of change of each landscape object. This explicit description could increase even more the ability to understanding the land use changes that are detectable in our remote sensing image databases.

ACKNOWLEDGMENTS

Gilberto Camara's work is partially funded by CNPq (grants PQ - 300557/19996-5 and 550250/2005-0) and FAPESP (grant 04/11012-0). Marcelino Silva's work is supported by UERN and funded by CAPES.

REFERENCES

AKSOY, S., KOPERSKI, K., TUSK, C. and MARCHISIO, G., 2004, Interactive Training of Advanced Classifiers for Mining Remote Sensing Image Archives. In *ACM International Conference on Knowledge Discovery and Data Mining*, (Seattle, WA: ACM). pp. 773-782.

ALMEIDA, A. L. O. D. and CAMPARI, J. S., 1995, *Sustainable Settlement in the Brazilian Amazon.* (New York: The World Bank).

ALVES, D., ESCADA, M. I. S., PEREIRA, J. L. G. and LINHARES, C. D. A., 2003, Land use intensification and abandonment in Rondônia, Brazilian Amazônia. *International Journal of Remote Sensing,* **24,** 899-903.

BATISTELLA, M., ROBESON, S. and MORAN, E., 2003, Settlement design, forest fragmentation, and landscape change in Rondonia, Amazonia. *Photogrammetric Engineering and Remote Sensing,* **69,** 805-812.

BECKER, B., 2004, *Amazônia - Geopolítica na Virada do III Milênio (Amazonia - Geopolitics on the Verge of the Third Millenium).* (Rio de Janeiro: Garamond).

BINS, L., FONSECA, L. and ERTHAL, G., 1996, Satellite Imagery Segmentation: a region growing approach. In *VIII Brazilian Symposium on Remote Sensing*, (São José dos Campos, BR: INPE). pp. 677-680.

CÂMARA, G., EGENHOFER, M., FONSECA, F. and MONTEIRO, A. M., 2001, What´s In An Image? In *Spatial Information Theory: Foundations of Geographic Information Science. International Conference, COSIT 2001.*, D. Montello (Ed.) (Santa Barbara, CA.: Springer), pp. 474-487.

CÂMARA, G., SOUZA, R., FREITAS, U. and GARRIDO, J., 1996, SPRING: Integrating Remote Sensing and GIS with Object-Oriented Data Modelling. *Computers and Graphics,* **15,** 13-22.

CAMPARI, J. S., 2002, Challenging the Turnover Hypothesis of Amazon Deforestation: Evidence from Colonization Projects in Brazil. Report (Austin: The University of Texas).

CHEN, Y., WANG, J. Z. and KROVETZ, R., 2003, CLUE: Cluster-based Retrieval of Images by Unsupervised Learning. In *Seventh International Symposium on Signal Processing and its Applications*, (Paris: IEEE).

DATCU, M., DASCHIEL, H., PELIZZARI, A., QUARTULLI, M., GALOPPO, A., COLAPICCHIONI, A., PASTORI, M., SEIDEL, K. and MARCHETTI, P. G., 2003, Information Mining in Remote Sensing Image Archives - Part A: System Concepts. *IEEE Trans. on Geoscience and Remote Sensing,* **41**.

ESCADA, M. I. S., MONTEIRO, A. M., AGUIAR, A. P., CARNEIRO, T. and CAMARA, G., 2005, Análise de padrões e processos de ocupação para a construção de modelos na Amazônia (Analysis of land use patterns and processes for the construction of models in Amazonia: Experiments in Rondonia). In *XII Brazilian Symposium on Remote Sensing*, (Goiania, Brazil: SELPER). pp. 2973-2983.

FONSECA, F., EGENHOFER, M., AGOURIS, P. and CÂMARA, G., 2002, Using Ontologies for Integrated Geographic Information Systems. *Transactions in GIS,* **6,** 231-257.

FORMAN, R. T. T., 1995, *Land Mosaics: The Ecology of Landscapes and Regions.* (Cambridge: Cambridge University Press).

GENELETTI, D. and GORTE, B. G. H., 2003, A method for object-oriented land cover classification combining Landsat TM data and aerial photographs. *International Journal of Remote Sensing,* **24,** 1273-1286.

INPE, 2005, Monitoramento da Floresta Amazônica Brasileira por Satélite (Monitoring the Brazilian Amazon Forest by Satellite). Available online at www.obt.inpe.br/prodes (acessed 19/09/2005).

LAMBIN, E. F., GEIST, H. J. and LEPERS, E., 2003, Dynamics of land-use and land-cover change in Tropical Regions. *Annual Review of Environment and Resources,* **28,** 205-241.

MACDONALD, J., 2002, The Earth Observation Business and the Forces that Impact It. In *Earth Observation Business Network 2002*, (Vancouver, CA: MacDonald Dettwiler).

MCGARIGAL, K., 2002, Landscape pattern metrics. In *Encyclopedia of Environmentrics*, A. H. El-Shaarawi and W. W. Piegorsch (Ed.) (Sussex, England: John Wiley & Sons), pp. 1135-1142.

MCGARIGAL, K. and MARKS, B., 1995, FRAGSTATS: spatial pattern analysis program for quantifying landscape structure. Report (Washington, DC: USDA Forestry Service Technical Report PNW-351).

MEINEL, G. and NEUBERT, M., 2004, A comparison of segmentation programs for high resolution remote sensing data. *International Archives of Photogrammetry and Remote Sensing,* **XXXV,** 1097-1105.

MERTENS, B. and LAMBIN, E., 1997, Spatial Modeling of Deforestation in Southern Cameroon: Spatial Disaggregation of Diverse Deforestation Processes. *Applied Geography,* **17,** 143--162.

QUINLAN, R., 1993, *C4.5: Programs for Machine Learning.* (San Francisco: Morgan Kaufmann).

RUSHING, J., RAMACHANDRAN, R., NAIR, U. J., GRAVES, S. J., WELCH, R. and LIN, A., 2005, ADaM: A Data Mining Toolkit for Scientists and Engineers. *Computers and Geosciences, In Press, Available online 11 January 2005*.

SHIMABUKURO, Y., BATISTA, G., MELLO, E., MOREIRA, J. and DUARTE, V., 1998, Using shade fraction image segmentation to evaluate deforestation in Landsat Thematic Mapper images of the Amazon region. *International Journal of Remote Sensing,* **19,** 535-541.

SILVA, M. P. S., CÂMARA, G., SOUZA, R. C. M., VALERIANO, D. M. and ESCADA, M. I. S., 2005, Mining Patterns of Change in Remote Sensing Image Databases. In *The Fifth IEEE International Conference on Data Mining*, (Houston.

TURNER, M. G., 1989, Landscape Ecology: The effect of Pattern on Process. *Annual Review of Ecology and Systematics,* **20,** 171-197.

WITTEN, I. H. and FRANK, H., 1999, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* (San Francisco: Morgan Kaufmann).

ZUCKER, S. W., 1976, Region growing: childhood and adolescence. *Computer Graphics and Image Processing,* **15,** 382-399.

APÊNDICE B

ARTIGO ACEITO EM CONFERÊNCIA INTERNACIONAL: THE FIFTH IEEE INTERNATIONAL CONFERENCE ON DATA MINING, 2005

Mining Patterns of Change in Remote Sensing Image Databases

Marcelino Pereira S. Silva[1,2], Gilberto Câmara[2], Ricardo Cartaxo M. Souza[2], Dalton M. Valeriano[2], Maria Isabel S. Escada[2]

[1]*UERN – Rio Grande do Norte State University, BR 110, Km 48, 59610-090, Mossoró, RN, Brazil*

[2]*INPE – National Institute for Space Research, P.O. Box 515, 12201-097, São José dos Campos, SP, Brazil*

{mpss, gilberto, cartaxo, isabel}@dpi.inpe.br, dalton@ltid.inpe.br

Abstract

Remote sensing image databases are the fastest growing archives of spatial information. However, we still have a limited capacity for extracting information from large remote sensing image databases. There are currently very few techniques for image data mining and information extraction in large image data sets, and thus we are failing to exploit our large remote sensing data archives. This paper proposes a methodology to provide guidance for mining remote sensing image databases. The basic idea is to use domain concepts to build generic description of patterns in remote sensing images, and then use structural approaches to identify such patterns in images. We illustrate our proposal with a case study for detecting land use patterns in Amazonia from INPE's remote sensing image database.

1. Introduction

Remote sensing satellites are currently the most significant source of new data about our planet, and remote sensing image databases are the fastest growing archives of spatial

information. The variety of spatial and spectral resolutions for remote sensing images ranges from IKONOS 1-meter panchromatic images to the next generation of polarimetric radar imagery satellites. Given the widespread availability of remotely sensed data, many government and private institutions have built large remote sensing image archives. The US National Satellite Land Remote Sensing Data Archive, managed by USGS EROS Data Center, holds 1,400 TB of satellite data collected over a 40 year period, and satellites such as NASA's Terra and Aqua generate an additional 3 Terabytes of imagery daily. Brazil's National Institute for Space Research (INPE) has more than 130 Terabytes of image datasets, covering 30 years of remote sensing activities, which are available on a database with free on-line access for Brazilian researchers. Strategic information from these remote sensing images is strongly demanded in many areas, including government (e.g., security and social purposes), economy (crop forecasting), and hydrology (water resources monitoring).

The first operational remote sensing satellite (LANDSAT-1) was launched in 1972, since then there has been a large worldwide experience in data gathering, processing and analysis of remotely sensed data. However, we still have a limited capacity for extracting information from large remote sensing image databases. Currently, most image processing techniques are designed to operate on a single image, and there are few algorithms and techniques for handling multi-temporal images [1]. This situation has lead to a "knowledge gap" in the process of deriving information from images and digital maps. This "knowledge gap" has arisen because there are currently very few techniques for image data mining and information extraction in large image data sets, and thus we are failing to exploit our large remote sensing data archives.

Although there has been a large research effort in content-based image retrieval (CBIR) techniques [2-6], the specific problem of mining remote sensing image databases has received much less attention. Proposals such as VISIMINE [7], ADAM [8] and KIM [9] are focused on clustering methods that operate on the feature space, the multi-dimensional space which is created by the different spectral bands of a remote sensing image. These techniques are useful for distinguishing spectral signatures of different land use types, such as finding areas which are classified as "lakes", "cities" or "forests".

However, in remote sensing image mining, one of the most important challenges is tracking patterns of land use change. A large remote sensing image database is a

collection of snapshots of landscapes, which provide us with a unique opportunity for understanding how, when, and where changes take place in our world. For example, INPE's image database covers a 30-year history of land use change in the Amazon tropical forest. Extensive fieldwork also indicates that the different actors involved in land use change (small-scale farmers, large plantations, cattle ranchers) can be distinguished by their different spatial patterns of land use [10]. Furthermore, these patterns evolve in time; new small farms will be created and large farms increase their agricultural area at the expense of the forest. In these and similar situations, patterns of land use change will have similar spectral signatures and image mining techniques based on clustering in the feature space will not be able to distinguish between them.

Therefore, tracking the temporal evolution of patterns in remote sensing imagery requires methods that are different from standard content-based image retrieval (CBIR) systems. A typical CBIR system uses a query image as the source and images in the database as targets, and query results are a set of images sorted by feature similarities with respect to the source [11]. When searching for patterns in remote sensing image databases, a different approach is necessary. Instead of similarity searches between image pairs, a system for mining remote sensing image databases must be able to do similarity searches between patterns found in different images. Therefore, mining remote sensing image databases is searching for patterns of change, not searching for internal content.

Our approach differs from previous work in the literature for content-based image retrieval. Schober et al [12] present a system that provides an automated keyword annotation for images, which assigns descriptive contents to objects in the image. Wang et al [5] describe an architecture that identifies object boundaries in a query image using segmentation, associates these objects to ontological concepts using neural networks, and uses these concepts to obtain a description for the image. These approaches aim at obtaining an adequate description of a single image, and are not adequate for solving the challenge of mining patterns in large remote sensing image databases, where the aim is to find similar patterns over significant temporal periods. We believe that by focusing on specific domain concepts for remote sensing data, it is possible to obtain significant results in mining land use patterns in large image databases.

Given this perspective, this paper proposes a methodology for mining patterns of change in remote sensing image databases. The basic idea is to use domain concepts to build

generic descriptions of patterns in remote sensing images, and then use structural approaches to identify these patterns in the image database. Our approach is motivated by earlier works by our research group on using ontologies for integrated GIS [13], and on ontological characterization of remote sensing imagery [14].

In what follows, we discuss patterns of change in remote sensing images and propose a methodology for mining land use patterns in remote sensing image databases. We illustrate our proposal with a case study for detecting land use patterns in Amazonia from INPE's remote sensing image database.

2. Patterns of change in remote sensing image databases

Given a large remote sensing image database, researchers would like to explore the database with questions such as: *What are the different land use patterns present in the database? When did a certain land use pattern emerge? What are the dominant land use patterns for each region? How do patterns emerge and change over time?* The answer to these and similar questions requires the availability of data mining techniques which are able to perform similarity searches between patterns found in different images. We propose to approach this problem by using spatial patterns as a means of describing relevant semantic features of an image.

Our primary consideration is that the instruments onboard remote sensing satellites capture energy at different parts of the electromagnetic spectrum, which is then converted into digital imagery. These instruments are not designed for a specific application, but are a compromise between sensor technology and requirements from different user communities. As a result, remote sensing images have a structural description which is independent of the application domain that a scientist employs to extract information. We distinguish between the image domain and the application domain, as shown in Figure 1:

- *Spatial Patterns* – the geometric structures that can be extracted from the images using techniques for feature extraction, segmentation, and image classification. They must be identified and labeled according to a typology which expresses their semantics. Examples of such patterns include *corridor-like regions* and *regular-shaped polygons* representing patterns of the mined data.

- *Application Concepts* – the different classes of spatial objects, which are associated to a specific domain. For example, in deforestation assessments, concepts include *large-scale agriculture, small-scale agriculture, cattle ranching* and *wood logging*.

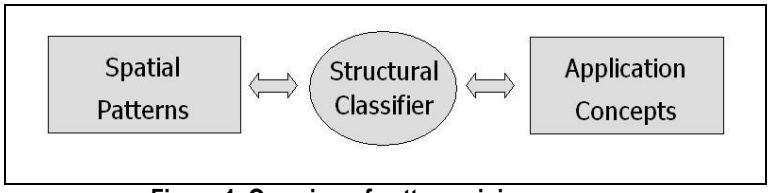


Figure 1. Overview of pattern mining process

To associate structures found in the image to concepts in the application, we need a *structural classifier*, which is able to relate the same structures to different application domains. This strategy differs from most remote sensing image database mining systems, such as KIM [9] and VISIMINE [7], which implicitly assume that there is one "best fit" for associating semantic concepts in the user domains to image-derived structures. Our view is that different *structural classifiers* will produce different associations between spatial patterns and the user domain concepts, and that each association is valid within a given application context. In other words, there are many ways to bridge the "sensory gap" and we should not search for a "best fit". For each type of application, there will be an appropriate structural classifier.

In what follows, we describe our proposed methodology for image mining, and apply it to the problem of mining patterns in INPE's remote sensing image database. In this paper, the application domain is concerned with describing land use change in tropical forests using remote sensing satellites.

3. A methodology for mining land use patterns on remote sensing images

We propose a methodology for image mining in large remote sensing databases using the idea application-dependent structural classifier, as outlined above. The methodology consists of three steps:

- Definition of a *spatial pattern typology* according to the user's application domain (Figure 2).
- Building a reference set of *spatial patterns*. This reference set is built using a prototypical set of images. *Landscape objects* are identified and labelled: the

identification employs image segmentation and the labeling is performed according to the *spatial pattern typology* (Figure 3).

- Mining the database using a *structural classifier* (guided by the *application concepts* of the domain), matching the reference set of *spatial patterns* to the *landscape objects* identified in images, thus revealing the *spatial configurations* present in each image (Figure 5).

3.1. Defining a spatial pattern typology

The first phase of the methodology calls for the definition of a spatial pattern typology which is associated to a given application domain. In order to illustrate our proposal, we will use a typology defined for mapping different types of land use change in tropical forests.

When using remote sensing images for understanding the forces driving changes in tropical forests, the assumption is that the expression of change is captured by changes in land use [15]. Extensive fieldwork also indicates that the different actors involved in land use change (small-scale farmers, large plantations, cattle ranchers) can be distinguished by their different patterns of land use [10]. Lambin et al. [10] propose a typology of the land use patterns in terms of deforestation processes (see Figure 2): *corridor* (commonly associated with riverside and roadside colonization), *diffuse* (generally related to smallholder subsistence agriculture), *fishbone* (typical of planned settlement schemes), and *geometric* (frequently linked to large-scale clearings for modern sector activities).

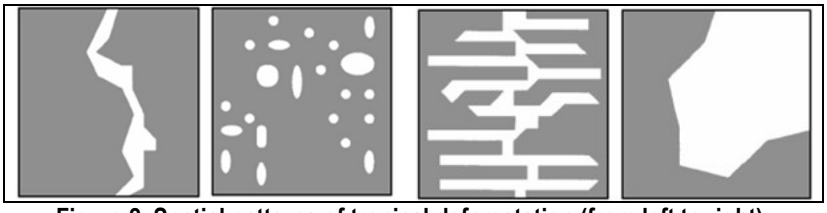


Figure 2. Spatial patterns of tropical deforestation (from left to right): corridor, diffuse, fishbone, and geometric (source: [10])

In this work, we will use the *spatial patterns typology* of Lambin et al., relating them to the structures of *landscape objects* in order to obtain the *spatial patterns*, through a *cognitive assessment* process, in which a human specialist associates *landscape objects* to *spatial patterns typology elements.*

3.2. Building a reference data set of spatial patterns

To represent the structures detected in remote sensing images, we introduce the concept of a *landscape object*. A landscape object is a structure detected in a remote sensing image by means of an image segmentation algorithm. *Landscape objects* can be associated to different types of spatial patterns.

To build a reference set of spatial patterns (Figure 3), we obtain a set of prototypical *landscape objects*, which are extracted from a set of sample images. We use segmentation algorithms to partition the image into regions which are spatially continuous, disjoint and homogenous [16]. Recent surveys [17] indicate that region-growing approaches [18] are well-suited for producing closed and homogenous regions. In our proposal, we have adopted the region-growing segmentation algorithm developed by INPE's [19], and grated in the SPRING software system [20], which is freely available on the Internet. This algorithm has been extensively validated for extracting land use patterns in tropical forests [21] and has been very favourably reviewed in a recent survey [17].

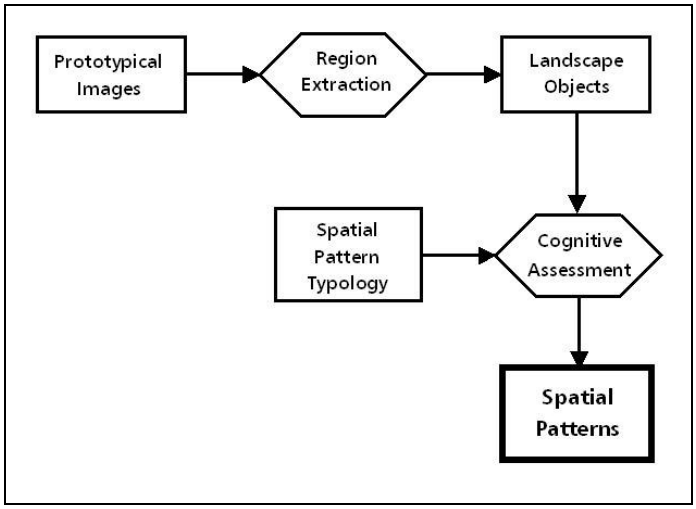


Figure 3. Building a reference set of spatial patterns

SPRING's region growing algorithm works as follows (Figure 4) [22]: (a) the image is first segmented into atomic cells of one or few pixels; (b) each segment is compared with its neighbors to determine if they are similar or not. If similar, they are merged and the mean gray level of the new segment is updated; (c) the segment continues growing by comparing it with all the neighbors until there is no remaining joinable region, at which point the segment is labeled as a completed region; and (d) the process moves to the next uncompleted cell,

repeating the entire sequence until all cells are labeled. The algorithm requires two parameters: (a) a similarity threshold value, and (b) an area threshold value.

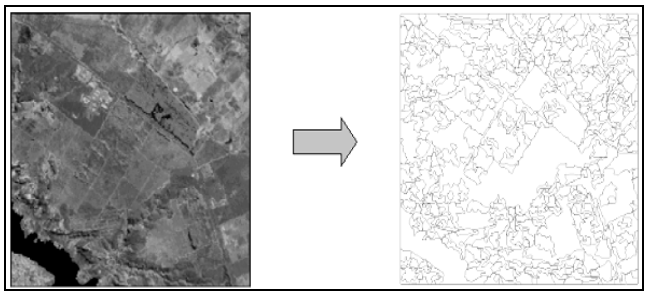


Figure 4. Example of a segmentation process

3.3. Mining the database using a structural classifier

Once the reference set of *spatial patterns* is built, the next phase will use them to mine *spatial configurations* from image databases. The *structural classifier* enables the association between *landscape objects* extracted from images and the reference set of *spatial patterns* (Figure 5).

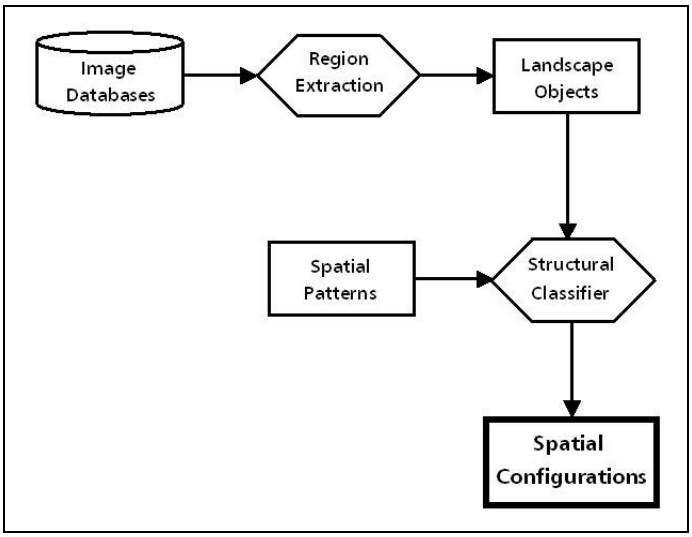


Figure 5. Obtaining spatial configurations

The *structural classifier* must be able to distinguish between different *spatial patterns*. It uses the C4.5 decision tree classifier [23], a classification method based on a decision tree. It predicts the value of a categorical attribute [24] based on non-categorical attributes. The

categorical attribute is the pattern type and the non-categorical attributes are a set of numerical attributes that characterize each pattern.

To select the attributes that distinguish the different types of land use patterns, we have used the concepts from *Landscape Ecology* [25]. Landscape ecology is based on the notion that environmental patterns strongly influence ecological processes. One of the key components of landscape ecology theory is the definition of metrics that characterize geometric and spatial properties of categorical map patterns [26]. The pattern metrics used in landscape ecology include metrics of spatial configuration that operate at the patch level. Patches form the building blocks for categorical maps and within-patch heterogeneity is ignored. Patch metrics refer to the spatial character and arrangement, position, or orientation of patches within the landscape. We have used the pattern metrics proposed by the FRAGSTATS (Spatial Pattern Analysis Program for Categorical Maps) software [27], that include:

- *Perimeter (m) and area (ha).*
- *Para* (perimeter-area ratio): a measure of shape complexity.
- *Shape* (shape index): patch perimeter divided by the minimum perimeter possible for a maximally compact patch of the corresponding patch area.
- *Frac* (fractal dimension index): two times the logarithm of patch perimeter (m) divided by the logarithm of patch area (m^2).
- *Circle* (related circumscribing circle): 1 minus patch area (m^2) divided by the area (m^2) of the smallest circumscribing circle.
- *Contig* (contiguity index): equals the average contiguity value for the cells in a patch.

The landscape ecology metrics are fed into the C4.5 classification algorithm to distinguish the different types of spatial patterns. After this classifier is properly trained, it can be used to label the landscape objects found in other images. Therefore, for each image in the database, this procedure identifies the number and location of the different types of spatial patterns. We refer to a specific set of spatial patterns found in an image as a *spatial configuration*.

By identifying the *spatial configurations* of different images, the user will be able to evaluate the emergence and evolution of different types of change. Each spatial pattern is associated to a different type of land use change. Therefore, the comparison between spatial configurations of images in different locations and between spatial configurations of images

at the same location in different times will allow new insights into the processes and actors that bring about change.

4. Case study: image mining for deforestation patterns

Brazil is facing a difficult challenge: controlling deforestation on Amazon rain forest, which covers about 40% of its territory. The causes of deforestation include economic, social and political factors and the current pace of land use change is substantial, with an average of 25,000 km^2 of forest being cleared every year. That situation demands fast and effective actions for reducing this pace of devastation. In order to monitor the extremely fast process of land use change in Amazonia, it is very important that INPE be able to use its huge data archive to the maximum extent possible. Given this motivation, we have used the methodology proposed above to achieve a better understanding of the processes of land use change in Amazonia.

We developed a case study using Landsat TM images (225/64, 226/64, 226/65, 225/65) of 1997, 2000, 2001, 2002 and 2003, which cover the region of São Félix do Xingu in the state of Pará. The images and deforestation data were provided by PRODES Project [28]. The *application concepts* for this task are guided by the land use change domain in tropical forests (Table 1).

Table 1. Land use change in tropical forests

Landscape object	Land use change
Corridor pattern	Roadside colonization Riverside deforestation
Diffuse pattern	Smallholder agriculture Small deforestation increments
Geometric pattern	Large farms

4.1. Building spatial patterns

According to the proposed methodology, *landscape objects* were extracted from prototypical images. Then, a human specialist, through *cognitive assessment,* obtained *spatial patterns* based on the *spatial patterns typology* of tropical deforestation (Figure 2). *Spatial patterns* are presented in Figure 6.

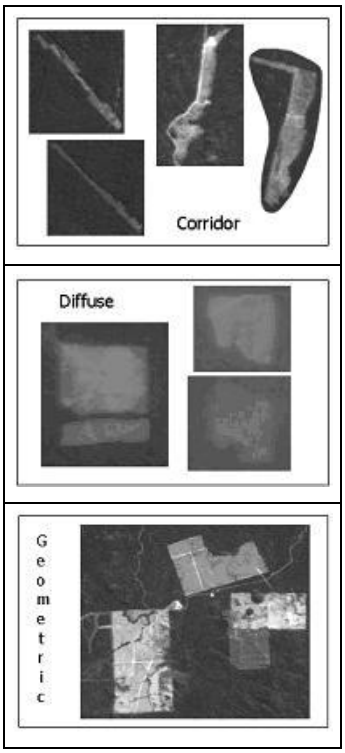


Figure 6. Spatial patterns representing corridor, diffuse and geometric patterns

4.2. Obtaining spatial configurations

The *structural classifier*, using the *spatial patterns*, extracted *spatial configurations* from the set of images just mentioned. Some results are presented below.

In a first case, we wanted to answer the following question: *"What's the behavior of large farmers in São Félix do Xingu during this period (1997-2003)? Is the area of new large farms increasing?"* Observing the evolution of the corresponding spatial configuration (geometric patterns) in Figure 7, it was possible to conclude that *"in 2000, this kind of deforestation reached a peak of 55,000 ha, but decreased in the following years. In 2003, the deforestation area associated to large farms decreased to 29,000 ha. This indicates that large farms are reducing their contribution to deforestation"*.

We posed a second question: *"What's the distribution of smallholder agriculture and small deforestation increments in São Félix do Xingu area during the years 1997-2003?"*. Observing Figure 8, we concluded that *"the distribution of this land use pattern in this period was mainly concentrated in the northeast and southeast of this area"*.

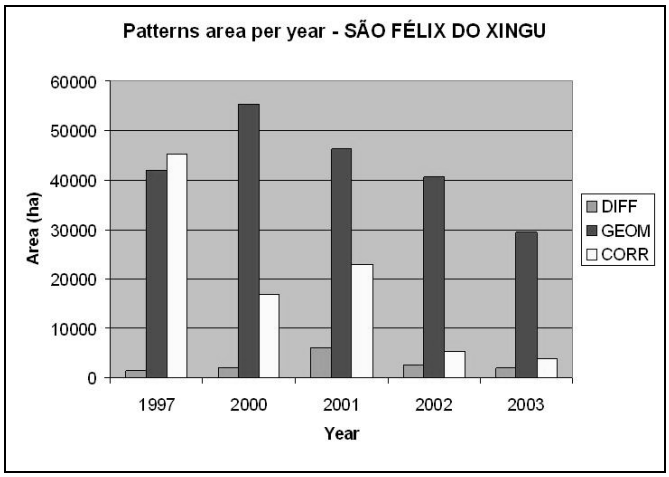


Figure 7. Large farms dynamic in São Félix do Xingu

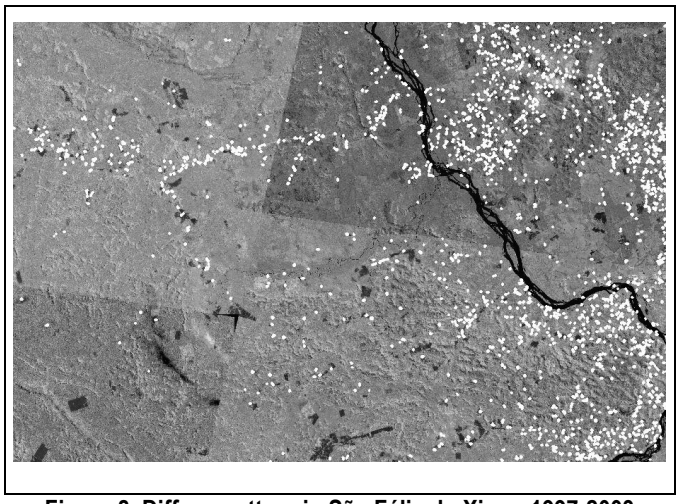


Figure 8. Diffuse pattern in São Félix do Xingu 1997-2003

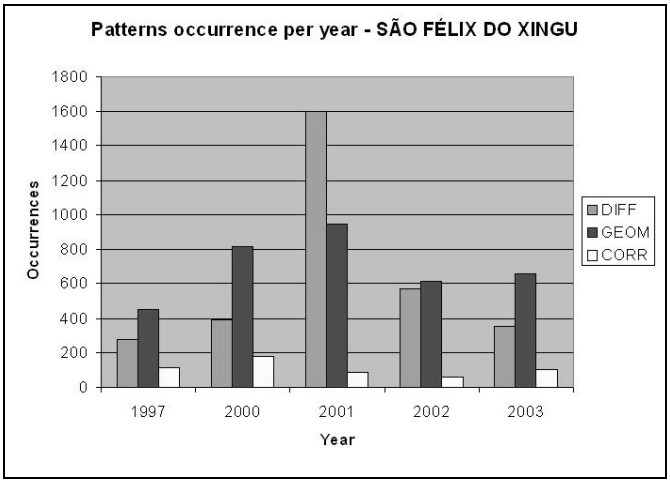


Figure 9. Diffuse patterns in São Félix do Xingu

The next question is: "*In São Félix do Xingu region, is there any dominant land use change pattern?"* Observing Figure 9, we concluded that: *"Diffuse pattern represented 61% of total occurrences of land use changes in 2001, indicating an increase in smallholder agriculture / small increments in deforested areas in that year".*

5. Conclusions

The methodology for mining patterns of change on remote sensing image databases proposed in this paper supports the extraction of spatial configurations and spatial patterns from these datasets. This methodology has been developed as an answer to the problem of searching for land use change patterns in remote sensing images. We consider that the proposed methodology can assist the environmental community to respond to the challenge of understanding and modeling land use change in a rapidly changing world. It also represents an alternative for making a more effective use of the large remote sensing image databases available in agencies such as USGS, ESA and INPE. Using satellite images and concepts of landscape ecology, the methodology provides a way to identify deforestation patterns in a complex domain, the Amazon forest. This approach bridges the gap between huge image databases and distinct domains (e.g. crop forecasting, deforestation).

Concepts and techniques of data mining, digital image processing and landscape ecology were used in the methodology to achieve good results during pattern detection. Images of distinct dates enabled the detection of pattern changes, which are extremely valuable when assessing, managing or preventing deforestation processes. Presented results revealed that the methodology is an important contribution to increase quality and speed of remote sensing image knowledge extraction.

Further experiments are necessary to calibrate distinct aspects of the methodology, such as image and structural patterns selection, segmentation and classification parameters, and mining algorithm aspects. Experimental evidences show that qualitative description of deforestation patterns are scale dependent: aspects like area and perimeter are relevant to identify land use change pattern. More specific deforestation patterns must enhance the detection and analysis of such processes, including specific deforestation actor definitions, enabling specialists to perform more accurate and faster tasks using specialized application concepts of their domains.

Acknowledgements

Contributions of Antônio Miguel V. Monteiro are gratefully acknowledged. Marcelino Silva would like to thank UERN and CAPES for supporting his work. Gilberto Câmara has been partially supported by CNPq and FAPESP.

References

[1] Datcu, M., et al., Information Mining in Remote Sensing Image Archives - Part A: System Concepts. IEEE Trans. on Geoscience and Remote Sensing, 2003. 41(2923--2936).

[2] Smeulders, A.W.M., et al., Content-Based Image Retrieval at the End of the Early Years. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000. 22(12): p. 1349 - 1380.

[3] Rui, Y., T.S. Huang, and S.-F. Chang, Image retrieval: current techniques, promising directions and open issues. Journal of Visual Communication and Image Representation, 1999. 10: p. 39--62.

[4] Wang, J.Z., J. Li, and G. Wiederhold, SIMPLIcity: Semantics-sensitive Integrated Matching for Picture LIbraries. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2001. 23(9): p. 947-963.

[5] Wang, L., L. Khan, and C. Breen. Object Boundary Detection for Ontology-based Image Classification. in Third International Workshop on Multimedia Data Mining. 2002. Edmonton, Alberta, Canada: ACM.

[6] Chen, Y. and J.Z. Wang, Image Categorization by Learning and Reasoning with Regions. Journal of Machine Learning Research, 2004. 5: p. 913-939.

[7] Aksoy, S., et al. Interactive Training of Advanced Classifiers for Mining Remote Sensing Image Archives. in ACM International Conference on Knowledge Discovery and Data Mining. 2004. Seattle, WA: ACM.

[8] Rushing, J., et al., ADaM: A Data Mining Toolkit for Scientists and Engineers. Computers and Geosciences, In Press, Available online 11 January 2005, 2005.

[9] Schröder, M., et al., Interactive Learning and Probabilistic Retrieval in Remote Sensing Image Archives. IEEE Trans. on Geoscience and Remote Sensing, 2000. 23(9): p. 2288--2298.

[10] Lambin, E.F., H.J. Geist, and E. Lepers, Dynamics of land-use and land-cover change in Tropical Regions. Annual Review of Environment and Resources, 2003. 28: p. 205-241.

[11] Chen, Y., J.Z. Wang, and R. Krovetz. CLUE: Cluster-based Retrieval of Images by Unsupervised Learning. in Seventh International Symposium on Signal Processing and its Applications. 2003. Paris: IEEE.

[12] Schober, J.-P., T. Hermes, and O. Herzog. Content-based Image Retrieval by Ontology-based Object Recognition. in KI-2004 Workshop on Applications of Description Logics (ADL-2004). 2004. Ulm, Germany.

[13] Fonseca, F., et al., Using Ontologies for Integrated Geographic Information Systems. Transactions in GIS, 2002. 6(3): p. 231-257.

[14] Câmara, G., et al., What´s In An Image?, in Spatial Information Theory: Foundations of Geographic Information Science. International Conference, COSIT 2001., D. Montello, Editor. 2001, Springer: Santa Barbara, CA. p. 474-487.

[15] Alves, D., et al., Land use intensification and abandonment in Rondônia, Brazilian Amazônia. International Journal of Remote Sensing, 2003. 24(4): p. 899-903.

[16] Pekkarinen, A., A method for the segmentation of very high spatial resolution images of forested landscapes. International Journal of Remote Sensing, 2002. 23(14): p. 2817-2836.

[17] Meinel, G. and M. Neubert, A comparison of segmentation programs for high resolution remote sensing data. International Archives of Photogrammetry and Remote Sensing, 2004. XXXV(Part B): p. 1097-1105.

[18] Zucker, S.W., Region growing: childhood and adolescence. Computer Graphics and Image Processing, 1976. 15: p. 382-399.

[19] Bins, L., L. Fonseca, and G. Erthal. Satellite Imagery Segmentation: a region growing approach. in VIII Brazilian Symposium on Remote Sensing. 1996. São José dos Campos, BR: INPE.

[20] Câmara, G., et al., SPRING: Integrating Remote Sensing and GIS with Object-Oriented Data Modelling. Computers and Graphics, 1996. 15(6): p. 13-22.

[21] Shimabukuro, Y., et al., Using shade fraction image segmentation to evaluate deforestation in Landsat Thematic Mapper images of the Amazon region. International Journal of Remote Sensing, 1998. 19(3): p. 535-541.

[22] Bins, L., et al. Satellite Imagery Segmentation: a Region Growing Approach. in VIII Brazilian Symposium on Remote Sensing. 1996.

[23] Quinlan, R., C4.5: Programs for Machine Learning. 1993, San Francisco: Morgan Kaufmann.

[24] Witten, I.H. and H. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. 1999, San Francisco: Morgan Kaufmann.

[24] Quinlan, R., C4.5: Programs for Machine Learning. 1993, San Francisco: Morgan Kaufmann.

[25] Turner, M.G., Landscape Ecology: The effect of Pattern on Process. Annual Review of Ecology and Systematics, 1989. 20: p. 171-197.

[26] McGarigal, K., Landscape pattern metrics, in Encyclopedia of Environmentrics, A.H. El-Shaarawi and W.W. Piegorsch, Editors. 2002, John Wiley & Sons: Sussex, England. p. 1135-1142.

[27] McGarigal, K. and B. Marks, FRAGSTATS: spatial pattern analysis program for quantifying landscape structure. 1995, USDA Forestry Service Technical Report PNW-351: Washington, DC.

[28] INPE, National Institute for Space Research: Prodes Project - Brazilian Amazon Forest Monitoring using Satellites. 2005. URL: http://www.obt.inpe.br/prodes/