MINISTÉRIO DA CIÊNCIA E TECNOLOGIA
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

O Uso de Séries Temporais e Mineração de Dados no Mapeamento de Cobertura do Solo e seus Padrões

Alana Kasahara Neves

Monografia da disciplina de Introdução ao Geoprocessamento (SER-300), ministrada pelo Dr. Antônio Miguel Vieira Monteiro.

INPE
São José dos Campos
2015

Resumo

Ainda não há entendimento completo sobre a dinâmica da evolução da paisagem na região Amazônica. Isto ocorre, em grande parte, devido à grande heterogeneidade de uso e ocupação que a Amazônia sofreu desde a implantação dos antigos projetos de colonização e dos novos projetos de infra-estrutura do governo federal. A aquisição de dados de sensores remotos é uma prática importante para auxiliar na resolução dessa problemática. Sensores remotos, como o MODIS (Moderate Resolution Imaging Spectroradiometer), têm sido responsáveis pela construção de séries temporais de longo prazo. Uma das técnicas utilizadas para manipular a grande quantidade de observações existentes em séries temporais é a mineração de dados. Assim, o trabalho teve como objetivo realizar uma classificação automática para mapeamento da cobertura do solo e seus padrões de arranjo espacial utilizando dados de séries temporais de EVI (Enhanced Vegetation Index) do MODIS. Foram extraídos 79 atributos das séries temporais MODIS para a órbita-ponto 227-068 (cena Landsat TM) para os anos de 2008 e 2010. Os atributos foram utilizados para gerar um modelo de classificação através do algoritmo Random Forests e as classes de interesse foram Área Agrícola, Pastagem e Floresta. Na segunda parte do trabalho, foram extraídas quatro métricas de paisagem em espaços celulares 10km x 10km contendo as classificações de Pastagem e Área Agrícola. Fez-se a classificação dos padrões de arranjo espacial na área de estudo. As classificações de cobertura da terra geradas foram bastante parecidas com os dados de referência, o mapeamento do TerraClass. Os padrões de arranjo espacial são os relacionados com polígonos contínuos, evidenciando a forte atividade agropecuária na área de estudo. Em estudos futuros, pretende-se aumentar a quantidade de classes de interesse e analisar anos mais espaçados.

Palavras chave: Mineração de dados, séries temporais, métricas da paisagem.

Sumário

1. Introdução

1.1. A Cobertura do Solo na Amazônia

A história de ocupação mais recente da Amazônia, a partir da década de 50, caracterizou-se pela expansão da fronteira agrícola, que resultou em um ritmo acelerado e extenso de transformações. Este período foi marcado por altas e continuadas taxas de desflorestamento em alguns estados da Amazônia Legal, principalmente nas áreas localizadas no chamado "arco do desmatamento" (BECKER, 1990, 2009).

O processo de ocupação foi induzido pelo governo através de diversas ações, como a abertura de estradas e o financiamento de grandes projetos agropecuários e de exploração mineral, também ocorrendo o processo da ocupação espontânea, decorrente dos grandes fluxos migratórios ocorridos durante as décadas de 70 e 80. Atualmente são encontradas grandes áreas de pastagem, agricultura, reflorestamento e floresta secundária, estando grande parte das florestas primárias limitadas às unidades de conservação (BECKER, 2009).

Conceitualmente, a evolução do desmatamento da Amazônia se inicia pelo corte seletivo da madeira de valor econômico, seguido pela queima e a implantação de agricultura ou pecuária. No entanto, atualmente acredita-se que este conceito seja uma simplificação do processo de desmatamento, pois não considera a diversidade e a complexidade dos sistemas biofísicos e humanos da Amazônia. Estudos indicam que a dinâmica de uso na Amazônia pode variar entre os diferentes estados e entre as diversas regiões de um mesmo estado (ESCADA, 2003).

Devido à sua complexidade, ainda não há entendimento completo sobre a dinâmica da evolução da paisagem na região Amazônica. Isto ocorre, em grande parte, devido à grande heterogeneidade de uso e ocupação que a Amazônia sofreu desde a implantação dos antigos projetos de colonização e dos novos projetos de infra-estrutura do governo federal. Para suprir a necessidade de entendimento desse fenômeno, o Centro Regional da Amazônia (INPE/CRA) produziu e disponibilizou em parceria com a EMBRAPA dados a respeito da cobertura da terra em toda a Amazônia Legal Brasileira, conhecido como projeto TerraClass – Mapeamento do Uso e Cobertura da terra nas Áreas Desflorestadas na Amazônia Legal (COUTINHO et al., 2013). O TerraClass apresenta à sociedade, de forma espacial e numérica, quais as principais

atividades atuais responsáveis pelo fenômeno do desflorestamento e já foi executado para o ano de 2008, 2010 e 2012.

Apesar de atingir o objetivo ao qual se propõe, a maior parte da interpretação e classificação realizada no projeto TerraClass é feita de forma visual e manual. Atualmente já existem alguns esforços na tentativa de automatizar sua metodologia e reduzir o tempo de interpretação.

1.2. Mineração de Dados e Séries Temporais

A aquisição de dados por sensoriamento remoto é uma prática de crescente importância e, cada vez mais, vem se tornando fundamental para o conhecimento de fenômenos terrestres. O estudo desses fenômenos somente por observações *in situ* necessitaria de uma dispendiosa quantidade de recursos (tempo, dinheiro e recursos humanos). Além disso, para que os dados coletados tenham utilidade e possam, de fato, gerar conhecimento, é necessário que eles sejam transformados em informação. A transformação de dados em informação é um processo que pode se tornar trabalhoso, impraticável manualmente, devido à grande quantidade de dados (CAMILO; DA SILVA, 2009).

Sensores remotos, como o MODIS (Moderate Resolution Imaging Spectroradiometer), têm sido responsáveis pela construção de séries temporais de longo prazo. O MODIS possui índices de vegetação como produtos, capazes de fornecer comparações temporais e espaciais das condições da vegetação global. São dois os produtos de índice de vegetação disponíveis: NDVI (Normalized Difference Vegetation Index) e EVI (Enhanced Vegetation Index). O primeiro é mais sensível à presença de pigmentos, como a clorofila, enquanto o segundo está mais relacionado com as variações na estrutura do dossel, como índice de área foliar (IAF), tipo de dossel, fisionomia vegetal e estrutura da copa (HUETE et al., 2002). Por esse motivo, o estudo de séries temporais de EVI permite a obtenção de inferências acerca da cobertura do solo.

Uma das técnicas utilizadas para manipular a grande quantidade de observações existentes em séries temporais é a mineração de dados. A mineração de dados consiste em uma ferramenta de auxílio que permite gerar informação a partir de uma grande quantidade de dados, através do processo de descobrimento de novas correlações,

padrões e tendências nos dados, utilizando tecnologias de reconhecimento de padrões e técnicas matemáticas e estatísticas (LAROSE, 2014).

2. Objetivos

Utilizar séries temporais MODIS (EVI) para:

► Identificar e mapear classes de interesse na cobertura do solo (classificação automática);

► Analisar os padrões de arranjo espacial resultantes da classificação.

3. Materiais e Métodos

3.1. Área de Estudo

A área de estudo (Figura 1) escolhida para o trabalho foi a órbita-ponto 227-068 do sensor TM do satélite Landsat 5. Essa cena localiza-se ao norte do estado do Mato Grosso (MT) e abrange partes de oito (8) municípios: Juara, Nova Canaã do Norte, Itaúba, Tabaporã, Porto dos Gauchos, Itanhangá, Ipiranga do Norte e Sinop. A cena está contida no chamado Arco do Desflorestamento faz parte da fronteira agropecuária da região, sendo o Mato Grosso um dos três estados com maior área desmatada na Amazônia (MARGULIS, 2003).

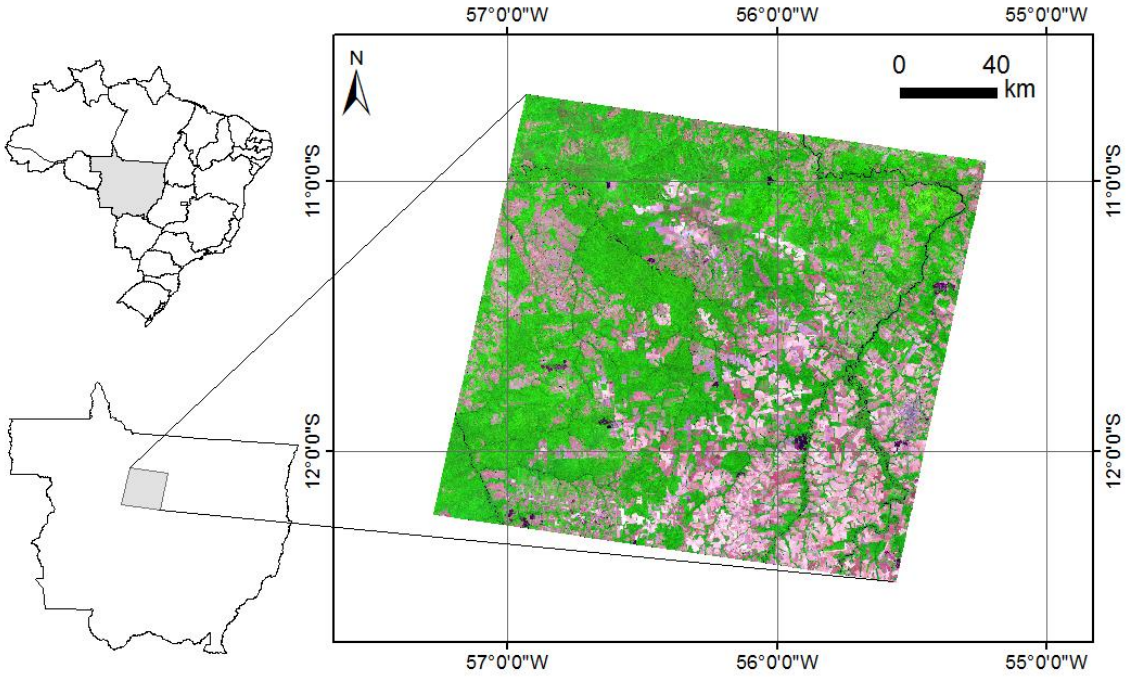


Figura 1. Área de estudo: órbita-ponto 227-068 do sensor TM (Landsat 5).

3.2. Classificação e validação

Primeiramente, foi realizada a extração de 79 atributos, pixel a pixel (Figura 2), do produto MOD13Q1 do MODIS, referente ao EVI. A série temporal consiste na variação dos valores de EVI ao longo do tempo para cada pixel. Esse produto possui uma resolução espacial de 250m e resolução temporal de 16 dias. O trabalho foi realizado para dois anos de estudo: 2008 e 2010. Devido à resolução temporal, são obtidos 23 momentos para cada ano. A extração de atributos foi realizada no GeoDMA 0.2.2 (Geographic Data Mining Analyst), plug-in do TerraView (Körting et al., 2008).

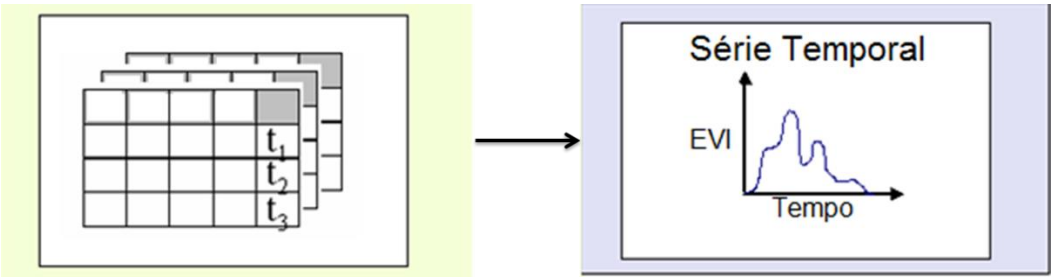


Figura 2. Extração de atributos pixel a pixel e composição de séries temporais (Adaptado de GOODALL, 2004).

Após a extração de atributos, realizou-se a classificação automática no software WEKA (HALL et al., 2009). Foi utilizado o algoritmo Random Forests, com duzentas (200) árvores de decisão, treinamento no ano de 2008 e avaliação no ano de 2010. Três classes de interesse foram discriminadas: Floresta, Pastagem e Área Agrícola.

O algoritmo Random Forests possui esse nome porque é composto por uma grande quantidade de árvores de decisão. Nesse algoritmo, os dados são particionados aleatoriamente em vários subconjuntos pela técnica de reamostragem Bootstrap (reamostragem com reposição, ou seja, alguns registros podem aparecer várias vezes no mesmo subconjunto enquanto outros não aparecem nenhuma vez). Cada subconjunto gera uma árvore de decisão e todas as árvores de decisão irão ter um voto com um determinado peso para contribuir na decisão da classe que será atribuída ao objeto de estudo (HAN et al., 2011).

Para validação, foram utilizados como referência os dados do TerraClass. A classificação gerada pelo TerraClass é feita a partir da intepretação de cenas do sensor TM do satélite Landsat 5. Para facilitar a comparação da classificação automática com a referência, apesar de os dados utilizados na classificação automática serem MODIS, a área de estudo é o box envolvente de uma cena Landsat. Como no projeto TerraClass existem vários tipos de pastos classificados, foi necessário fazer uma reclassificação, de modo que as classes Pasto Limpo e Pasto Sujo tornaram-se uma única classe chamada Pastagem, facilitando assim a comparação na validação.

Para testar a eficiência do modelo de classificação, existem medidas de avaliação e desempenho. Como medidas de desempenho da classificação, foram geradas Matrizes de Confusão e Índices Kappa para cada ano de estudo. A Matriz de Confusão mostra a taxa de acerto do modelo em predizer novas amostras das classes de interesse e o índice Kappa é um índice estatístico que diz quanto o modelo criado é melhor do que uma classificação meramente aleatória.

3.3. Padrões de Arranjo Espacial

O termo “paisagem” refere-se a um mosaico heterogêneo composto por padrões espaciais homogêneos, que podem ser relacionados com a dinâmica daquela paisagem. Entender os padrões de arranjo espacial existentes em uma área de estudo nos dá a possibilidade de extrair informações acerca da complexidade das mudanças ocorrendo

naquela região. A melhor maneira de identificar os padrões existentes é utilizar variáveis para a extração de informações, as chamadas métricas de paisagem (SAITO, 2011).

As métricas de paisagem são extraídas a partir da representação do espaço em unidades celulares. Foram utilizadas unidades celulares de tamanho 10 km x 10 km. A decisão do tamanho das células é feita de forma empírica de forma a evitar a perda de informação ou a mistura de diversas tipologias em uma única célular, conforme visto na figura 3.

No presente trabalho, foram utilizadas quatro métricas de paisagem (DAL'ASTA et al., 2013):

- Percentual de Paisagem (Percent Land);
- Densidade de manchas (Edge Density – ED);
- Dimensão fractal média da mancha (Mean Patch Fractal Dimension – MPFD);
- Índice de área ponderada pela forma média (Area Weight Mean Shape Index – AWSI).

As métricas de paisagem foram extraídas para cada célula contendo padrões classificados automaticamente (etapa anterior) de Pastagem e Área Agrícola (Figura 4) também pelo plug-in GeoDMA e suas descrições, fórmulas, intervalos e unidades podem ser vistas em Körting et al. (2008).

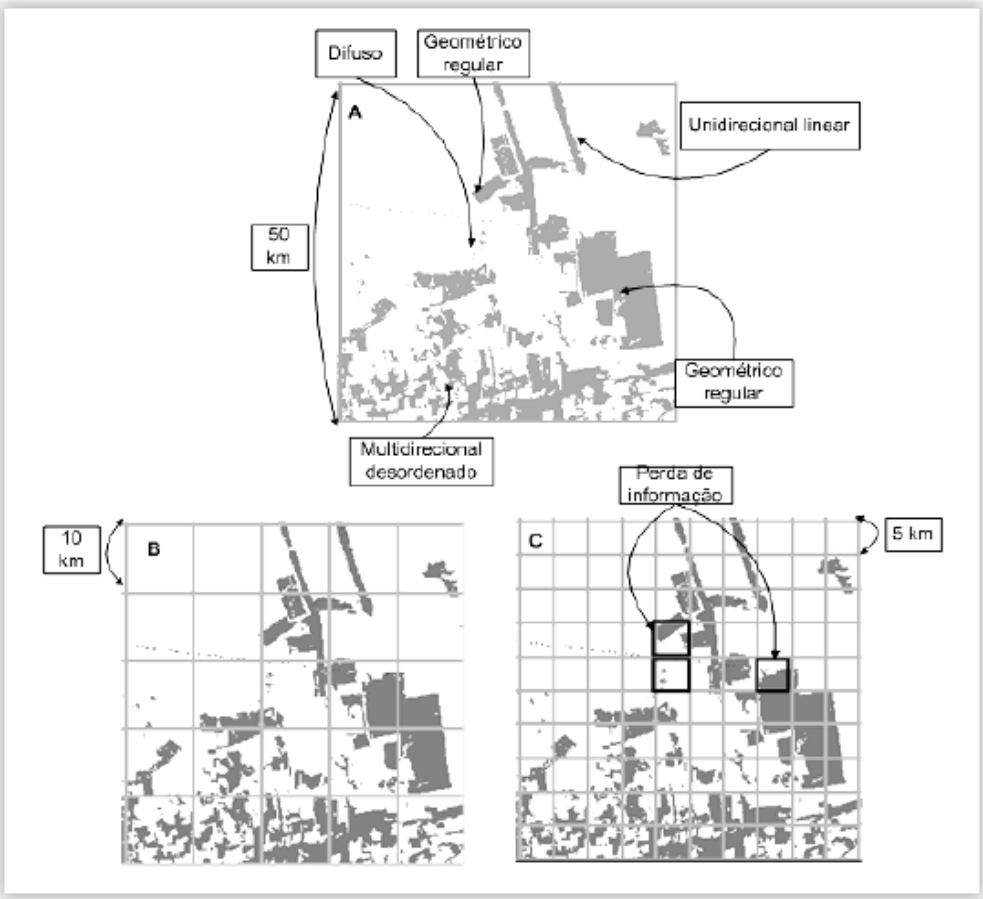


Figura 3. Escolha do tamanho das células. A) Mistura de diversos padrões. B) Boa discriminação dos padrões. C) Perda de informação (FONTE: SAITO, 2008).

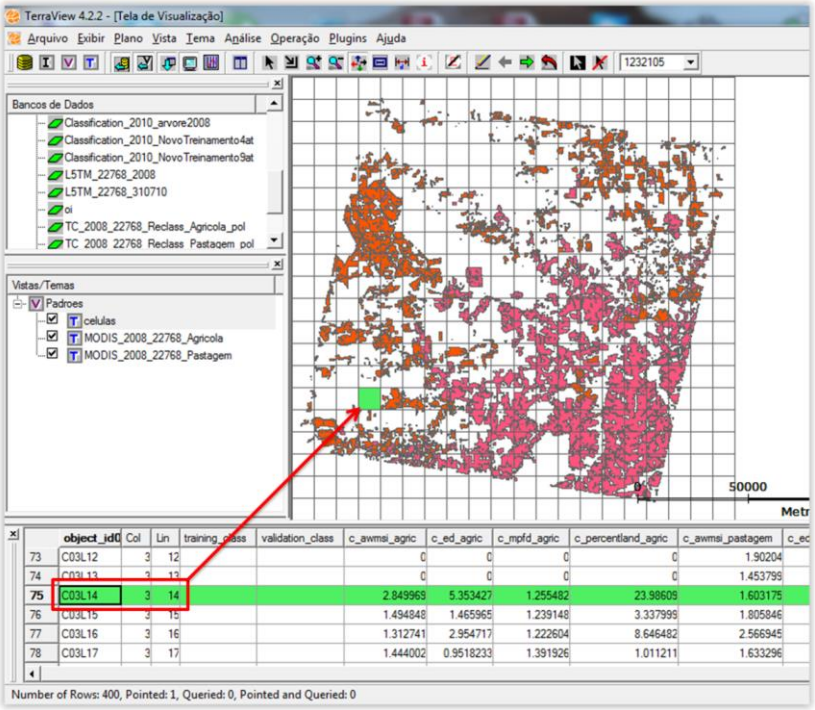
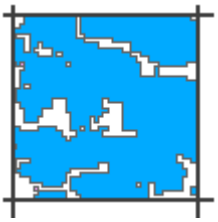
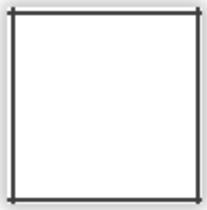


Figura 4. Extração das métricas de paisagem para cada célula, contendo Pastagem (em laranja) ou Área Agrícola (em rosa).

Após a extração das métricas da paisagem, foi feita a classificação dos espaços celulares de acordo com as tipologias da Tabela 1, baseadas em trabalhos anteriores (SAITO, 2008; DAL'ASTA ET AL., 2013; GAVLAK, 2011; ESCADA, 2003). A classificação foi feita utilizando o algoritmo C4.5.

Tabela 1. Tipologias de arranjo espacial

Padrão	Tipologia	Descrição
	Irregular contínuo	Segmentos e densidade de médios a grande, associados a formas não regulares e contínuas. Podem estar associadas a grandes áreas de pastagem e agricultura.
	Contínuo	Representam regiões com classes não agropecuárias, como floresta, área urbana, hidrografia, etc.
	Geométrico	Formatos geométricos regulares, com densidade baixa a média. Podem estar associados a talhões de agricultura mecanizada ou pecuária de grande porte.
	Geométrico contínuo	Segmentos e densidade de médios a grande, associados a formas geométricas regulares e contínuas. Podem estar associadas a grandes

		áreas de pastagem e latifúndios.
	Difuso	Manchas isoladas e pequenas. Podem estar associadas à agricultura itinerante ou a pequena produção.
	Misto	Formas irregulares associadas com formas geométricas: pequenas áreas agrícolas próximas à áreas de produção mecanizada ou pecuária de grande porte.

4. Resultados e Discussão

As classificações resultantes para a área de estudo nos anos de 2008 e 2010 ao lado de suas referências (classificação TerraClass) podem ser vistas nas figuras 5 e 6, respectivamente, nas quais as Pastagens estão representadas em laranja, as Áreas Agrícolas em rosa, a Floresta em verde e as regiões não classificadas (outras classes do TerraClass que não foram observadas, como Não Floresta, Área Urbana, Mosaicos de Ocupação, etc).

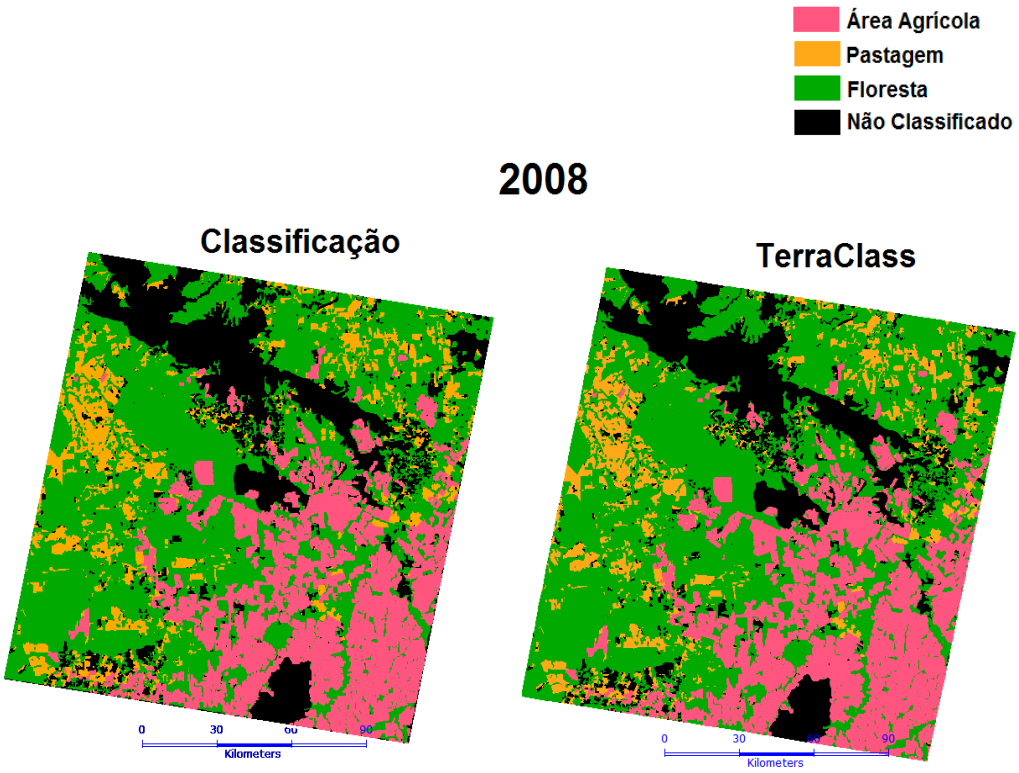


Figura 5. Classificação automática para 2008 e sua referência.

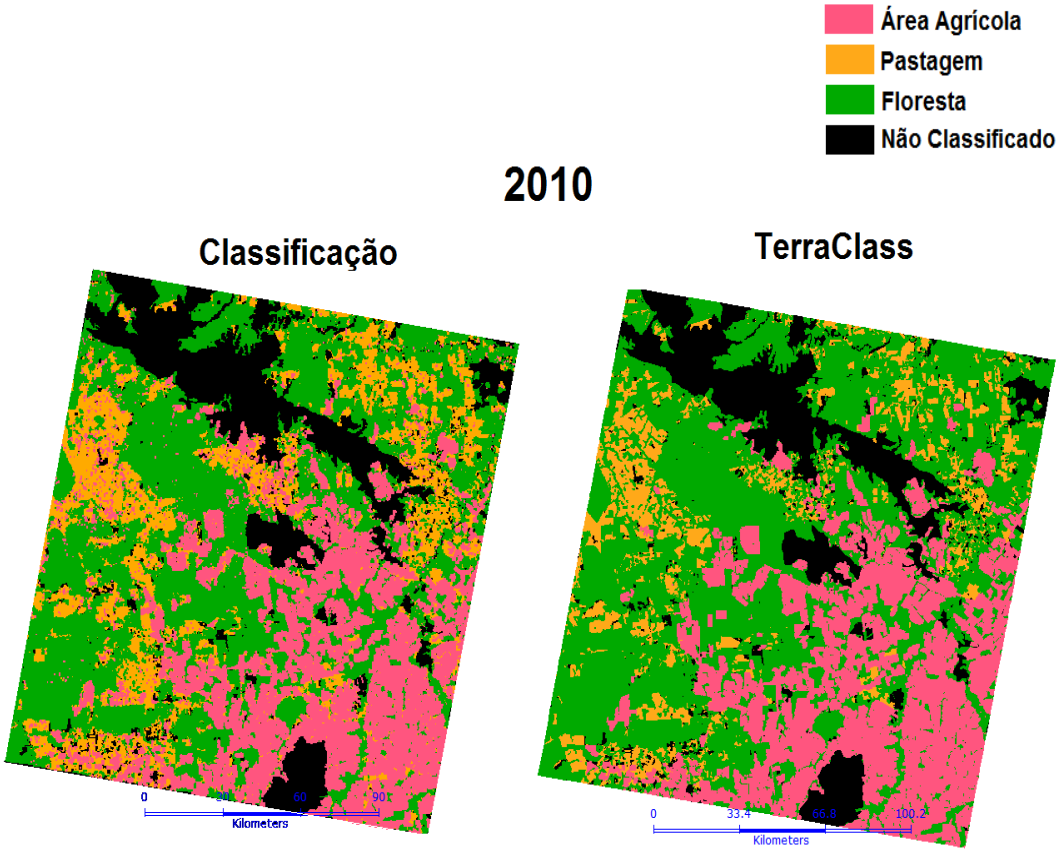


Figura 6. Classificação automática para 2010 e sua referência.

Como pôde ser observado, as classificações tiveram resultados bastante semelhantes com a referência. Entretanto, devido às diferenças de resoluções espaciais (Classificação automática: MODIS – 250m; Classificação de referência: TerraClass baseado em imagens do sensor TM do satélite Landsat – 30 m), que podem ser vistas na Figura 7, ocorrerão diferenças nas áreas classificadas pelas duas diferentes metodologias.

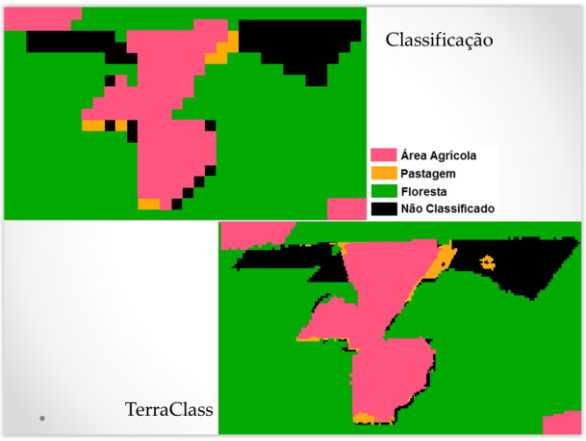


Figura 7. Diferenças de classificação devido às resoluções espaciais.

As matrizes de confusão para as classificações dos anos de 2008 e 2010 podem ser vistas, respectivamente, nas tabelas 2 e 3. Os acertos encontram-se na diagonal principal, ou seja, na classificação automática do primeiro ano, 85,46% do que foi classificado como floresta era realmente floresta, de acordo com o mapa de referência e assim por diante. No geral, foram classificados corretamente 86,08% dos pixels em 2008 e 78,89% em 2010. O índice Kappa para o ano de 2008 foi de 0,7912 e para 2010 foi de 0,6659.

Tabela 2. Matriz de Confusão para Classificação de 2008.

Classificação				
Floresta	**Pastagem**	**Área de Cultivo**		
16593 (85,46%)	1.678 (8,30%)	808 (4,50%)	**Floresta**	**Referência**
1819 (9,37%)	16.659 (83,36%)	823 (4,58%)	**Pastagem**	
1004 (5,17%)	1.889 (9,34%)	16.337 (90,92%)	**Área de Cultivo**	
19416 (100%)	20.226 (100%)	17.968 (100%)	**Total**	

Tabela 3. Matriz de Confusão para Classificação de 2010.

Classificação				
Floresta	**Pastagem**	**Área de Cultivo**		
147.736 (97,01%)	39.746 (36,77%)	9.026 (10,60%)	**Floresta**	**Referência**
2.350 (1,54%)	52.724 (48,78%)	3.975 (4,67%)	**Pastagem**	
2.210 (1,45%)	15.620 (14,45%)	72.144 (84,73%)	**Área de Cultivo**	
152.296 (100%)	108.090 (100%)	85.145 (100%)	**Total**	

O índice Kappa e o total de pixels classificados para 2010 foram mais baixos do que os valores encontrados em 2008. Isso ocorre o desenvolvimento do modelo para classificação ocorrer em 2008 e ser avaliado em 2010, portanto já era esperado que o acerto do modelo seria maior no conjunto de dados em que ele foi criado.

Em relação aos padrões de arranjo espacial, podemos ver na Figura 8 a árvore de decisão gerada pelo algoritmo C4.5.

```
c_ed_agric <=  0.7500352263
|   c_awmsi_pastagem >  1.7406580448  -> Geometrico_continuo
|   c_awmsi_pastagem <=  1.7406580448
|   |   c_awmsi_pastagem <=  0.0000000000  -> Continuo
|   |   c_awmsi_pastagem >  0.0000000000  -> Difuso
c_ed_agric >  0.7500352263
|   c_awmsi_pastagem <=  1.5678509474
|   |   c_awmsi_agric <=  2.4805719852  -> Geometrico
|   |   c_awmsi_agric >  2.4805719852  -> Irregular_continuo
|   c_awmsi_pastagem >  1.5678509474
|   |   c_awmsi_agric <=  1.5993269682  -> Geometrico
|   |   c_awmsi_agric >  1.5993269682  -> Misto
```

Figura 8. Árvore de Decisão para os Padrões de Arranjo Espacial

Nas figuras seguintes, temos os resultados da classificação para os padrões de arranjo espacial na cena de estudo. Nota-se que a classificação foi satisfatória em relação às pastagens e áreas agrícolas presentes na região, uma vez que as tipologias contínuas foram encontradas nas áreas de maior adensamento de polígonos nas células, os padrões difusos foram encontrados em áreas de polígonos pequenos, conforme era esperado na Tabela 1 da metodologia.

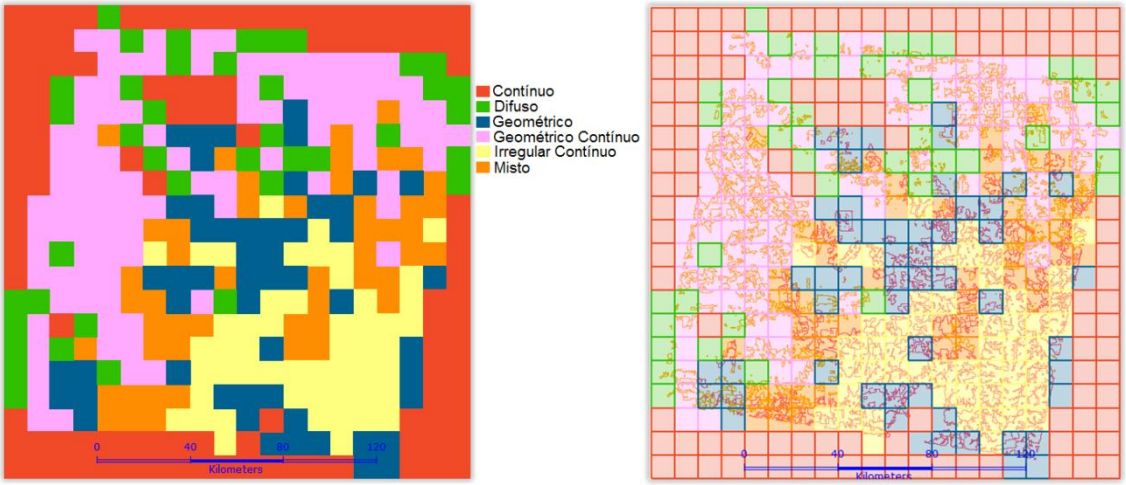


Figura 9. Classificação das Tipologias de Arranjo Espacial. A) Apenas as Tipologias. B) Tipologias e polígonos de Pastagem e Áreas Agrícolas embaixo.

Na tabela seguinte, temos discriminados quais as quantidades de células em cada tipologia, do total de 304 células (foram retiradas as células de borda na contagem).

Tabela 4. Células classificadas em cada tipologia.

Tipologia	Quantidade de células	% Total
Contínuo	13	4,28%
Difuso	40	13,16%
Geométrico	53	17,43%
Geométrico Contínuo	96	31,58%
Irregular contínuo	52	17,1%
Misto	50	16,45%
Total	304	100%

A maioria das células (31,58%) está associada à tipologia Geométrico Contínuo, indicando grande quantidade de áreas destinadas a grandes pastagens ou a grandes áreas agrícolas. Além disso, a menor quantidade de células (4,28%) está associada à tipologia Contínuo. Nota-se que esse é um resultado coerente, visto que a área de estudo no Mato Grosso consiste em uma região fortemente tomada por áreas de agricultura mecanizada e grandes pastagens, com poucas áreas de remanescente da floresta Amazônica.

5. Considerações Finais

A partir do trabalho realizado, pode-se confirmar a eficiência do algoritmo random forests na classificação de cobertura da terra, conforme já havia sido testado em Sato et al. (2013). As classes de estudo obtiveram uma discriminação satisfatória a partir da utilização de séries temporais de EVI do MODIS.

A área de estudo, no Mato Grosso, possui grande parte de sua área ocupada pela agropecuária. Pretende-se testar a metodologia para outras regiões da Amazônia e testar sua eficiência.

Como etapas futuras, pretende-se incorporar a classificação de novas classes relevantes de estudo, como a vegetação secundária. Além disso, também se pretende fazer a análise dos padrões de arranjo espacial para classificações de anos mais espaçados e, assim, construir trajetórias de cobertura da terra.

Referências Bibliográficas

BECKER, B. K. Amazônia. *Série Princípios.* São Paulo: Ática, 1990. 92p.

BECKER, B. K. Amazônia: Geopolítica na virada do III milênio. Rio de Janeiro: Garamond, 2009. 172p.

CAMILO, C. O.; DA SILVA, J. C. **Mineração de dados: Conceitos, Tarefas, Métodos e Ferramentas.** Relatório Técnico. Goiânia: UFG, 2009.

COUTINHO, A. C.; ALMEIDA, C.; VENTURIERI, A.; ESQUERDO, J. C. D. M.; SILVA, M. **Projeto TerraClass: Uso e cobertura da terra nas áreas desflorestadas na Amazônia Legal**. Brasília, DF: Embrapa; Belém: INPE, 2013.

DAL'ASTA, A. P.; ESCADA, M. I. S.; AMARAL, S.; MONTEIRO, A. M. V. Evolução do arranjo espacial urbano e das terras agrícolas no entorno de Santarém (Pará) no período de 1990 a 2010: Uma análise integrada baseada em sensoriamento remoto e espaços celulares. XVI Simpósio Brasileiro de Sensoriamento Remoto. **Anais...** Foz do Iguaçu, Paraná, 2013.

ESCADA, I. 2003. **Evolução de padrões de uso e cobertura da terra na região centro-norte de Rondônia**. Tese (Doutorado em Sensoriamento Remoto). INPE, São José dos Campos. 166p.

GAVLAK, AA (2011). **Padrões de mudança de cobertura da terra e dinâmica populacional no Distrito Florestal Sustentável da BR-163: população, espaço e ambiente.** 177 p. (sid.inpe.br/mtc-m19/2011/08.02.16.24-TDI). Dissertação (Mestrado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2011.

GOODALL, J. L.; MAIDMENT, D. R.; SORENSON, J. **Representation of Spatial and Temporal Data in ArcGIS.** AWRA GIS and Water Resources III Conference, Nashville, TN. 2004.

HALL M.; FRANK, E., HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. **The WEKA Data Mining Software: An Update**. SIGKDD Explorations, Volume 11, Issue 1. 2009.

HAN, J.; KAMBER, M.; PEI, J. **Data mining**: concepts and techniques. 3ed. San Francisco: Morgan Kaufmann Publishers, 2011.

HUETE, A.; DIDAN, K.; MIURA, T.; RODRIGUEZ, E. P.; GAO, X.; FERREIRA, L. G. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. **Remote Sensing of Environment**, v. 83, n. 1, p. 195-213, 2002.

KORTING, T. S.; FONSECA, L. M.; ESCADA, M. I. S.; SILVA, F. C.; SILVA, M. P. S. GeoDMA: a novel system for spatial data mining. **IEEE International Conference on Data Mining Workshops, Pisa, Italia**, 2008. **Anais…** Pisa, Italia, 2008.

LAROSE, D. T. **Discovering Knowledge in Data: An Introduction to Data Mining**. John Wiley and Sons, Inc, 2 ed. 2014.

MARGULIS, S. **Causas do Desmatamento da Amazônia Brasileira**. 1ed. Brasília: Banco Mundial, 2003.

SAITO, E. A. **Caracterização de trajetórias de padrões de ocupação humana na Amazônia Legal por meio de mineração de dados**. 2011. 160 f. Dissertação

(Mestrado em Sensoriamento Remoto) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos. 2011.