Análise e mineração de dados de sensores orbitais para acompanhamento de safras de cana-de-açúcar

Bruno F. Amaral[1], Daniel Y. Chino[1], Luciana A. S. Romani[2], Renata R. V. Gonçalves[3], Elaine P. M. de Sousa[1] , Agma J. M. Traina[1]

[1]Instituto de Ciências Matemáticas e de Computação – USP – São Carlos - Brasil

[2]Embrapa Informática Agropecuária – Campinas – SP – Brasil

[3]Cepagri – Universidade Estadual de Campinas – SP – Brasil

{brunoslash, chinodyt}@grad.icmc.usp.br, luciana@cnptia.embrapa.br, renata@cpa.unicamp.br, {parros, agma}@icmc.usp.br

Abstract. *Researches aiming greenhouse gases reduction have been motivated by the impact of extreme climate events around the world. In Brazil, sugar cane is the main source for ethanol production to replace fossil fuels. In this context, remote sensing imagery has been widely used to monitor sugar cane harvests and to support scientific research. In this paper, we propose a methodology based on data clustering to analyze NDVI time series obtained from AVHRR/NOAA satellites and monitor the growing cycles of sugar cane crops. The experiments show that our approach can identify areas with similar development patterns also considering different growing crops seasons.*

Resumo. *O impacto causado por eventos climáticos extremos em todo o mundo tem motivado pesquisas para redução de gases de efeito estufa. No Brasil, a cana-de-açúcar é a principal fonte para produção de etanol, como alternativa a combustíveis fósseis. Nesse contexto, dados de sensoriamento remoto têm sido utilizados para monitorar safras de cana-de-açúcar e apoiar pesquisas científicas. Neste trabalho, é proposta uma metodologia baseada em agrupamento de dados para analisar séries temporais de NDVI obtidas de satélites AVHRR/NOAA. Os experimentos mostram que a abordagem proposta permite identificar áreas com padrões de desenvolvimento similares, considerando também os diferentes ciclos de vida da cultura.*

1. Introdução

O aquecimento global e as mudanças climáticas são grandes desafios para os pesquisadores de diversas áreas do conhecimento. Apesar da certeza quanto à elevação da temperatura média do planeta, muitas pesquisas sobre os impactos no ecossistema terrestre e na agricultura vêm sendo desenvolvidas. No Brasil, uma das principais fontes de recursos vem da produção agrícola, como o biocombustível a partir da cana-de-açúcar, usado como fonte de energia renovável para substituir combustíveis fósseis. Desse modo, as pesquisas e estudos sobre o desenvolvimento das culturas agrícolas ao longo dos anos, de acordo com as mudanças no clima, são muito importantes. Os três cenários de mudanças climáticas (baixa, igual e alta emissão de carbono em relação aos dias atuais), analisados em função da emissão de gases de efeito estufa (GEE), têm mostrado impactos significativos na saúde, na agricultura, nos recursos hídricos, na biodiversidade e, particularmente, na incidência de eventos climáticos extremos. É

prevista a elevação nos níveis de evaporação e intensificação do ciclo hidrológico, decorrente da maior quantidade de energia disponível gerada pelo aquecimento, contribuindo para maior ocorrência de eventos extremos de precipitação, com chuvas mais intensas em determinadas áreas. O contrário também é observado, pois em algumas regiões as estiagens podem se tornar mais severas e prolongadas [IPCC 2007].

Nesse contexto, embora a cana-de-açúcar possa se beneficiar do aumento das temperaturas, um aumento do déficit hídrico em determinadas fases do desenvolvimento da cultura pode impactar sua expansão para áreas mais quentes. Portanto, estudos mais aprofundados e detalhados sobre a expansão e produtividade da cana-de-açúcar, bem como o desenvolvimento de novos cultivares, tornam-se fundamentais. Como a cana-de-açúcar é plantada em grandes extensões, é possível usar imagens de sensores remotos para auxiliar no monitoramento das plantações dessa cultura. Neste trabalho, imagens do satélite AVHRR/NOAA (*Advanced Very High Resolution Radiometer / National Oceanic and Atmospheric Administration*) são usadas para obtenção de índices, como o índice de vegetação extraído para cada pixel da imagem, gerando um grande volume de dados. Nesse cenário, o suporte computacional adequado para análise e extração de conhecimento desses dados se torna uma ferramenta muito útil.

Uma técnica importante para a análise de séries temporais de índices de vegetação visando descoberta e estudo de padrões no desenvolvimento de culturas agrícolas é a detecção de agrupamento (*clustering*) [Romani et al. 2011]. Neste trabalho, algoritmos de agrupamento são avaliados para auxiliar na visualização da expansão da cana-de-açúcar no estado de São Paulo, ao longo dos anos, de modo regional. Assim, este artigo apresenta a integração de tarefas desde extração de séries temporais a partir de índices das imagens de satélite até a identificação de agrupamento e apresentação espacial dos resultados em uma única ferramenta computacional.

2. Metodologia

Este trabalho propõe uma abordagem para analisar o desenvolvimento de culturas de cana-de-açúcar ao longo de safras por meio da utilização de técnicas de *clustering* aplicadas a séries temporais extraídas de imagens de satélite. O conjunto de dados é obtido a partir de sensores de satélites de baixa resolução, utilizando a ferramenta SatImagExplorer [Chino et al. 2010] para extração de índices vegetativos a partir de imagens de satélite. Na mesma ferramenta, foram implementados os métodos de *clustering* utilizados neste trabalho e aplicados para análise das séries temporais. As etapas principais do processo de análise dos dados são ilustradas na Figura 1.

Na primeira etapa, as imagens são obtidas por meio do sensor AVHRR do satélite NOAA. Este sensor tem sido muito usado em pesquisas relativas a ecossistemas e agricultura devido a um grande volume de dados disponíveis e por ter cobertura global, além da possibilidade de acesso gratuito aos dados. Na segunda etapa é realizada a combinação de canais desse sensor, o que fornece uma indicação da quantidade e estado da vegetação, como o NDVI (*Normalized Difference Vegetation Index*), um dos índices vegetativos utilizados no monitoramento de culturas por imagens de satélite. Um exemplo de cultura que pode ser monitorada por imagens de baixa resolução, como as do AVHRR/NOAA, é a cana-de-açúcar, por ocupar áreas vastas e bem próximas.

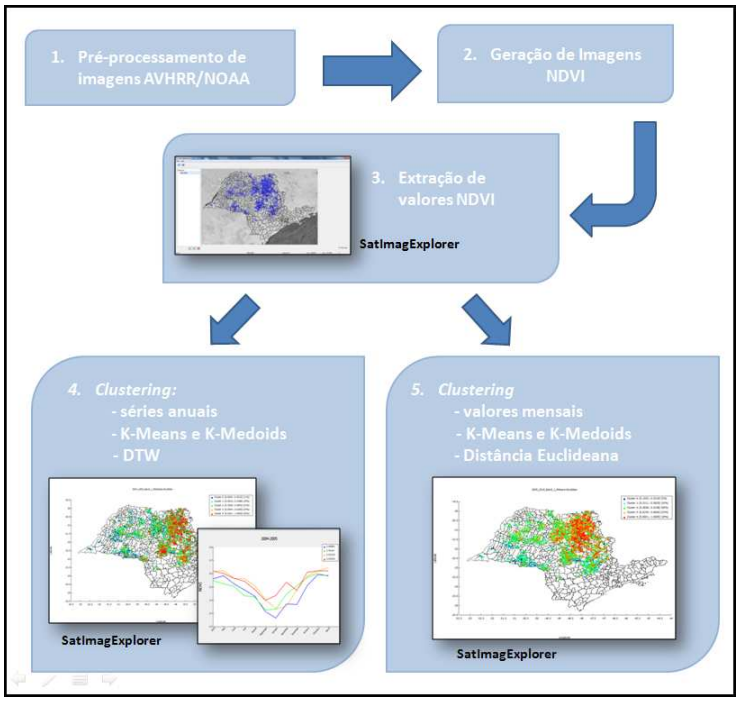


Figura 1. Metodologia utilizada neste trabalho.

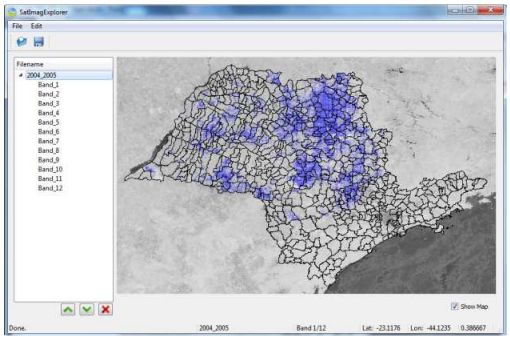


Figura 2. SatImagExplorer com imagem de São Paulo.

Na SatImagExplorer foram implementadas e incorporadas técnicas de *clustering*, incluindo suporte à descoberta de padrões nos dados extraídos de imagens NDVI. Assim, na etapa 3 a ferramenta recebe imagens de satélite como entrada e calcula índices e medidas de sequências, de modo automático, gerando séries temporais relativas a esses índices, de acordo com a área de interesse definida pelo especialista. Na Figura 2, é exibida a visualização dos valores de NDVI na SatImagExplorer. Nas etapas seguintes, o especialista pode aplicar os métodos de *clustering* que farão o agrupamento dos pontos da região do mapa, de acordo com os valores NDVI. Os métodos implementados são o *K-Means* e o *K-Medoids*, ambos descritos em [Han et al. 2001].

O *K-Means* é um algoritmo de *clustering* baseado em particionamento, em que cada objeto do conjunto de dados a ser agrupado pode pertencer a apenas um grupo (*cluster*). Inicialmente são escolhidos aleatoriamente os elementos representantes (*centroids*) dos *clusters*, e cada elemento da base é associado ao *cluster* cujo representante é o mais próximo. Utilizando o valor médio dos elementos de cada grupo, o algoritmo recalcula iterativamente os *centroids*, e redistribui todos os objetos, até que não haja mais alterações no agrupamento. Por ser baseado em valores médios, a presença de *outliers* e ruídos na base de dados pode influenciar na escolha dos *centroids*

e na formação dos *clusters*. O *K-Means* foi implementado na SatImagExplorer por ser um algoritmo rápido, simples e que demonstra bons resultados, além de ser muito conhecido e usado em diversas aplicações na área de mineração de dados.

O *K-Medoids* é um método de *clustering* semelhante ao *K-Means*, baseado em particionamento. Ele difere do anterior quanto à escolha do representante do *cluster*. No *K-Medoids*, ao invés de representar o grupo pela média de seus elementos, o algoritmo busca, como representante, o elemento mais central do *cluster*, chamado *medoid*. Dessa maneira, para cada *cluster*, todos os objetos pertencentes a ele são testados como candidatos a *medoid*, e apenas o que for o mais centralizado é escolhido. Tal característica torna o método menos sensível a ruídos e *outliers* no conjunto de dados de entrada, além de minimizar o efeito da escolha aleatória dos *medoids* iniciais. Entretanto, isso eleva significativamente seu tempo de execução.

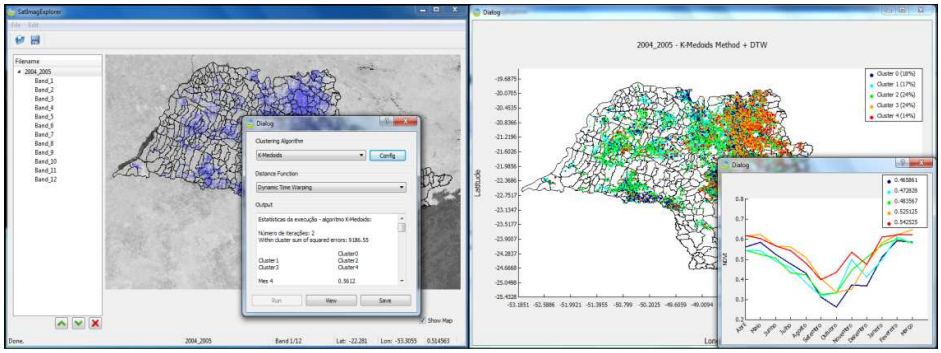


Figura 3. Aplicação dos métodos de *clustering* na SatImagExplorer.

Os métodos de *clustering* implementados utilizam métricas para calcular as distâncias entre os elementos do conjunto de dados. Neste trabalho, foram aplicadas duas abordagens na análise dos dados. Em uma delas, cada elemento do conjunto de dados é definido por apenas um valor de NDVI, referente ao índice de um mês de uma safra em uma determinada localização (pixel da imagem), visando análises mensais da região de interesse. Nesse caso, os métodos de *clustering* utilizam a distância Euclideana. A outra abordagem visa trabalhar com as séries de valores de NDVI, de modo que cada elemento do conjunto é definido por uma série de NDVI correspondente a uma ou várias safras de cana-de-açúcar. Nessa abordagem, foi utilizada a métrica DTW (*Dynamic Time Warping*) com ambas as técnicas de *clustering*, por ser uma função de distância muito aplicada na análise de séries temporais [Berndt and Clifford 1994]. A interface para o processo de *clustering* na ferramenta é ilustrada na Figura 3.

Em vários estudos [Ding et al. 2008], métricas simples como a distância Euclideana são suficientes para cálculo de similaridade em diversos domínios de dados. Entretanto, séries temporais podem possuir formatos ou perfis semelhantes, porém estar desalinhadas em relação à escala temporal. A DTW, por outro lado, realiza alinhamentos entre duas séries, reconhecendo a similaridade mesmo com deslocamentos de tempo. Na Seção 3, são descritos experimentos usando as duas abordagens.

3. Experimentos e Discussão

A metodologia descrita neste trabalho foi aplicada na análise de séries temporais de imagens de NDVI no período de 2001 a 2010 para o estado de São Paulo. Para selecionar os potenciais pixels de cana-de-açúcar foi utilizada uma máscara gerada pelo

projeto CANASAT/INPE[1] para a safra 2004/2005 como referência. Dois experimentos foram realizados:

1. Agrupamento de pixels mês a mês para cada ano-safra utilizando os algoritmos *K-Means* e *K-Medoids* com a função de distância Euclideana;
2. Agrupamento de pixels por década utilizando o algoritmo *K-Medoids* com a função de distância DTW.

Em ambos os experimentos os pixels foram separados em 5 *clusters*. No experimento 1, os pixels correspondem a um único valor mensal de NDVI e por isso são agrupados utilizando a distância Euclideana. No experimento 2, são utilizados todos os pixels da série temporal, isto é, 12 valores mensais para cada ano-safra, perfazendo um total de 120 valores de NDVI na série. Nesse experimento foi utilizada a função de distância DTW, que é mais utilizada para detecção de similaridade em séries temporais.

A fim de gerar um *baseline* para a comparação dos resultados gerados pelo modelo proposto de análise de agrupamento foram definidas, pelos especialistas, classes baseadas nos valores de NDVI considerando a safra 2004/2005, a saber:

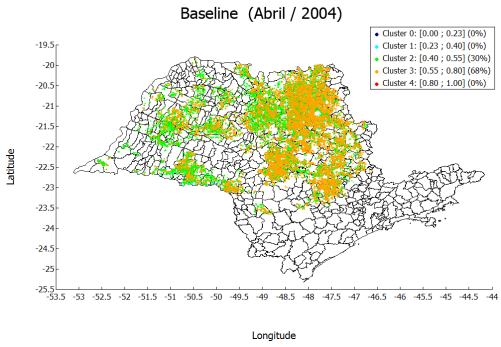
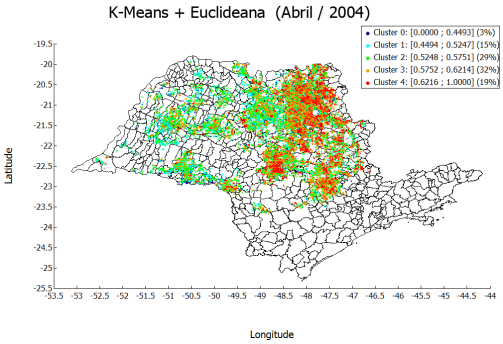
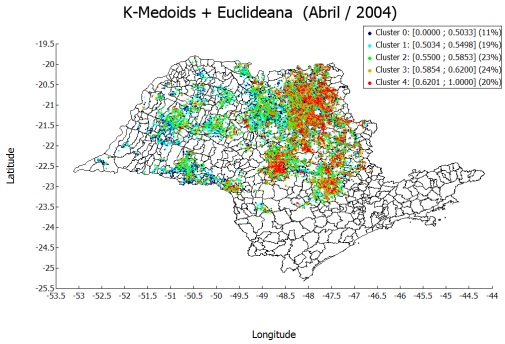
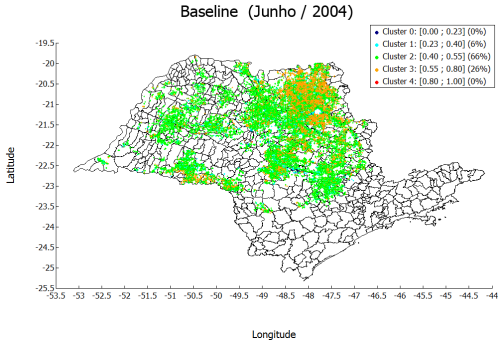
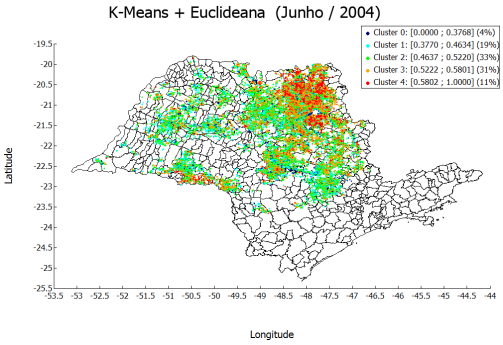
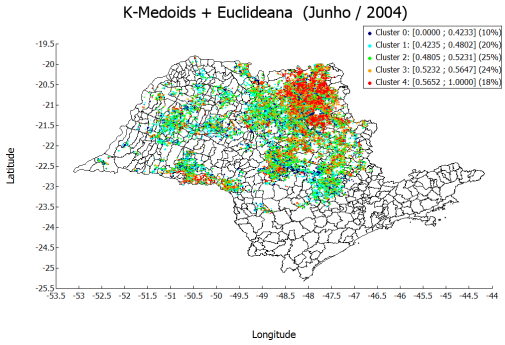
- *Cluster* 0 (azul escuro) = 0,0 a 0,23
- *Cluster* 1 (azul claro) = 0,23 a 0,40
- *Cluster* 2 (verde) = 0,40 a 0,55
- *Cluster* 3 (laranja) = 0,55 a 0,80
- *Cluster* 4 (vermelho) = 0,80 a 1,0

Foram geradas imagens mensais de acordo com esse *baseline*, como mostrado nas Figuras 4A, 4D, 4G, 4J e 4N. No primeiro experimento, ambos os métodos de agrupamento produziram resultados semelhantes, onde o *cluster* 4 (vermelho) indica os valores máximos de NDVI no mês correspondendo a áreas com maior biomassa e, consequentemente, maior produtividade. O *cluster* 0 (azul escuro) mostra os valores de NDVI mais baixos, correspondendo, provavelmente, a solo exposto.

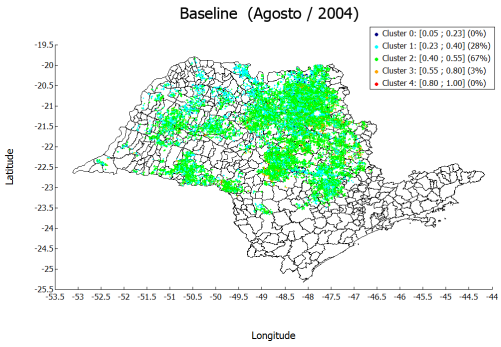
O mês de Abril corresponde ao ponto máximo do crescimento vegetativo da cana-de-açúcar. Na Figura 4A (*baseline*), os pixels aparecem em laranja com o valor máximo de NDVI, correspondendo às áreas em vermelho e laranja (acima de 0,58) geradas pelo *K-Means* e *K-Medoids*, nas Figuras 4B e 4C respectivamente. Logo os dois métodos apresentaram resultados similares à classificação proposta pelos especialistas.

Por outro lado, o mês de Outubro corresponde à época de colheita. Neste mês, os pixels em azul claro e escuro (com o valor mínimo de NDVI) na imagem *baseline* (Figura 4J) correspondem às áreas em verde, azul claro e escuro (abaixo de 0,37) geradas pelo *K-Means* e *K-Medoids*, nas Figuras 4L e 4M. Novamente, ambos os algoritmos mostraram-se similares aos resultados indicados pela classificação prévia. Análises similares podem ser realizadas para os demais meses ilustrados na Figura 4. Assim, aplicando os métodos de *clustering* na análise da safra de cana-de-açúcar (1 ano de abril a março), é possível acompanhar o ciclo da cultura ao longo dos meses, automaticamente. Com isso, demonstra-se que embora as técnicas usadas sejam bem conhecidas, a sua aplicação em séries geradas a partir de imagens de baixa resolução permite estudo da cultura em nível regional. Além disso, como não existe a predominância em todas as regiões de um ou dois agrupamentos ao longo dos meses, é possível observar que canas de ano e ano e meio são cultivadas em todo o estado.

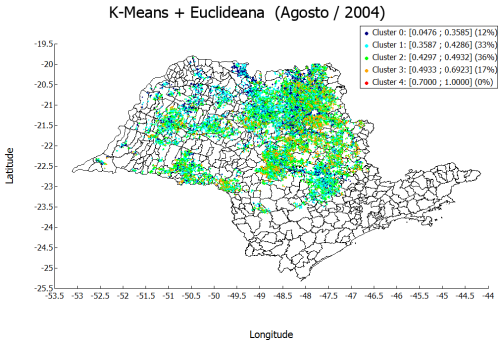
[1] www.dsr.inpe.br/laf/canasat

	***K-Means* + Distância Euclideana**	***K-Medoids* + Distância Euclideana**
Abril (2004/05)	 ***Baseline* (A)**	
	 B	 **C**
Junho (2004/05)	 ***Baseline* (D)**	
	 **E**	 **F**

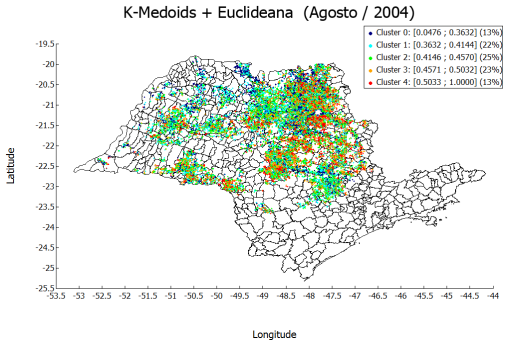
Agosto (2004/05)



***Baseline* (G)**

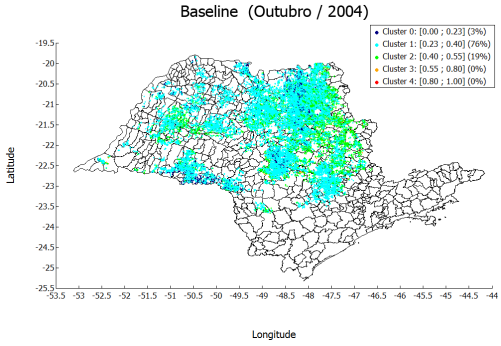


H

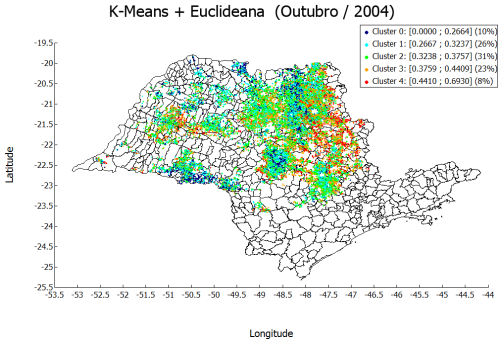


I

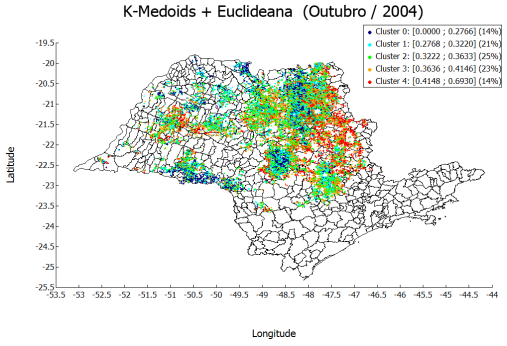
Outubro (2004/05)



***Baseline* (J)**



L



M

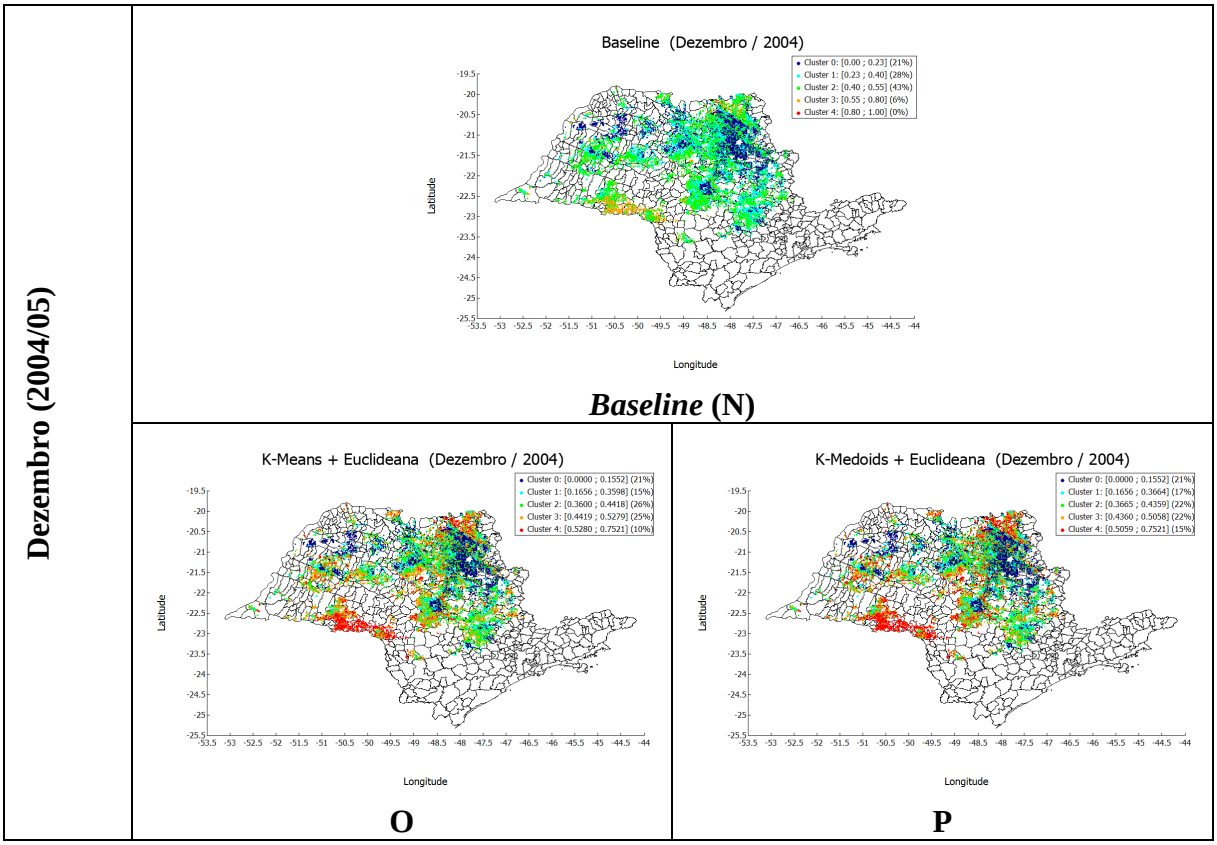


Figura 4. Mapas de São Paulo com 5 *clusters* (*baseline*, *K-Means* e *K-Medoids*).

No experimento 2, os resultados ilustrados na Figura 5 mostram que, embora tenha sido usada uma máscara para mapear a safra de cana-de-açúcar de 2004/2005, é possível diferenciar as regiões tradicionalmente produtoras (noroeste) das áreas de expansão (oeste) de cana-de-açúcar. A apresentação das diferentes áreas produtoras de cana-de-açúcar em um único mapa que representa a análise de uma década de imagens de NDVI permite que o especialista valide resultados relativos à produção de cana-de-açúcar. Além disso, tanto a técnica quanto a integração das tarefas em um único sistema tornam o trabalho do especialista mais rápido e eficiente, permitindo uma análise automática regional uma vez que são utilizadas imagens de alta resolução temporal. Na Figura 5A, os pixels do *cluster* 4 (vermelho), com 7%, apresentam um perfil diferente do padrão de NDVI da cana-de-açúcar, como pode ser visto na assinatura espectral gerada para o *medoid* do *cluster* 4 (vermelho), na Figura 5B. É possível que essa área corresponda a outro tipo de vegetação, como matas ou reflorestamento. Por outro lado, o *cluster* 3 (laranja), com 34%, representa a área de maior produção da cana-de-açúcar. O perfil de NDVI para esse *cluster*, ilustrado na Figura 5B, mostra uma tendência mais semelhante ao perfil padrão de NDVI, atingindo os maiores valores do índice.

Os *clusters* 1 e 2 (azul-claro e verde), com 22% e 26% respectivamente, indicam a área de expansão da cana-de-açúcar para o oeste do estado (Figura 5A). Os perfis de NDVI dos dois *clusters* não atingem valores muito expressivos, indicando que a produção pode não ser tão elevada quanto da região noroeste do estado. O *cluster* 0 (azul escuro), com 8%, aparece distribuído em todo o estado e de acordo com o seu perfil, corresponde a solo exposto ou mistura espectral.

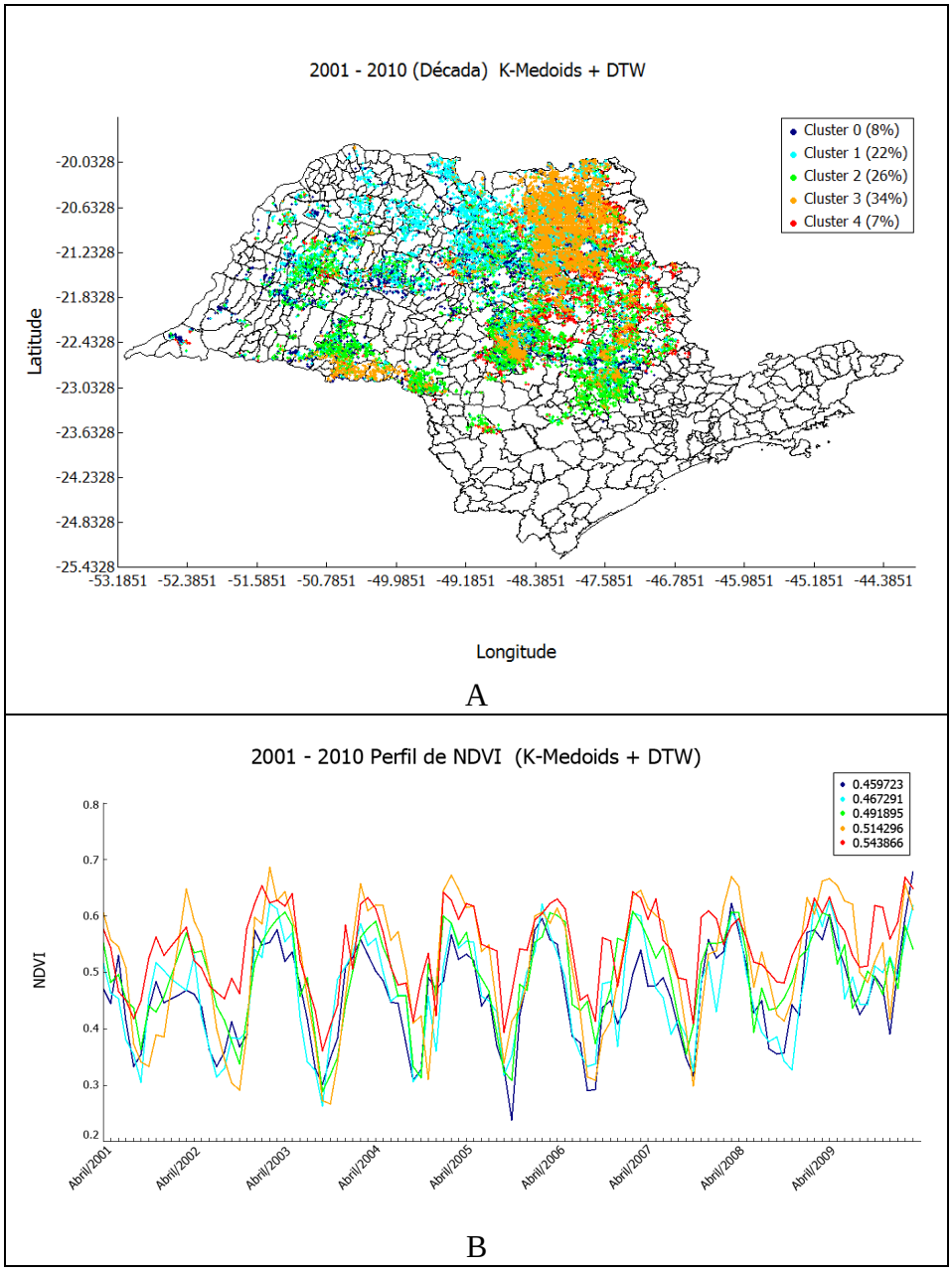


Figura 5. A. Mapa de São Paulo com 5 *clusters* (*K-Medoids* + DTW) no período de 2001 a 2010. B. Gráfico com a assinatura espectral (perfil de NDVI) dos 5 *clusters*.

A validação dos resultados foi efetuada de duas maneiras: pelo especialista e por meio da comparação com implementações disponíveis em outro software de mineração de dados, o Weka[2]. De acordo com os especialistas, os resultados são promissores, pois foi possível acompanhar a expansão da cana-de-açúcar. Os resultados da implementação no SatImagExplorer e no Weka foram similares. No entanto, o SatImagExplorer tem a vantagem de integrar todo o processo, desde a extração das séries de NDVI até a apresentação espacial do resultado da análise de agrupamento.

[2] http://www.cs.waikato.ac.nz/ml/weka/

4. Conclusões

Este trabalho propõe uma nova abordagem para avaliar a expansão da cana-de-açúcar, por meio de técnicas de mineração de dados e de séries temporais. Foram usadas técnicas de *clustering* tradicionais associadas a funções de distância para séries temporais. Todo o processo desde a extração dos pixels das imagens de satélite até a análise e visualização dos resultados foi desenvolvida em uma ferramenta (SatImagExplorer), tornando o processo mais fácil e ágil para o especialista.

Embora os métodos computacionais utilizados não sejam inovadores, a sua aplicação para apoiar o acompanhamento das safras de cana-de-açúcar, bem como averiguar sua expansão de maneira automática, é uma valiosa contribuição deste trabalho. Além disso, o potencial das técnicas utilizadas destaca-se ainda mais por terem sido aplicadas a imagens de satélite de baixa resolução espacial, o que dificulta a análise devido à possibilidade de mistura espectral. A realização das análises numa única ferramenta computacional como o SatImagExplorer, que incorpora funcionalidades desde a extração das séries a partir das imagens até a apresentação visual dos resultados, auxilia nas pesquisas sobre fontes renováveis de energia, como o etanol. O impacto deste tipo de trabalho torna-se ainda maior à medida que aumenta a necessidade por pesquisas para a redução de gases do efeito estufa, diante das recentes ocorrências de eventos extremos em diferentes localidades do planeta.

Agradecimentos

Os autores agradecem Santander, Fapesp-Microsoft Research, CNPq, Capes e Embrapa pelo apoio financeiro e Cepagri/Unicamp pelas imagens do AVHRR/NOAA.

Referências

Berndt, D. and Clifford, J. (1994) “Using dynamic time warping to find patterns in time series”. In: AAAI Workshop on Knowledge Discovery in Databases, p. 359-370, Seattle - Washington.

Chino, D. Y. T., Romani, L. A. S. e Traina, A. J. M. (2010) “Construindo séries temporais de imagens de satélite para sumarização de dados climáticos e monitoramento de safras agrícolas”. In: REIC, v. 10, p. 1-16.

Ding, H., Trajcevski, G., Scheuermann, P., Wang, X. and Keogh, E. (2008). “Querying and Mining of Time Series Data: Experimental Comparison of Representations and Distance Measures”. In: VLDB Endowment, p. 1542-1552.

IPCC (2007) “Climate change 2007: Fourth assessment report (AR4)”. In: Intergovernmental Panel on Climate Change.

Han, J., Kamber, M. and Tung, A. K. H. (2001) “Spatial Clustering Methods in Data Mining: A Survey”. In: Geographic Data Mining and Knowledge Discovery, Edited by H. J. Miller and J. Han, Taylor and Francis, p. 201-230.

Romani, L. A. S., Gonçalves, R. R. V., Amaral, B. F., Zullo Jr, J., Traina Jr, C., Sousa, E. P. e Traina, A. J. M. (2011) “Acompanhamento de safras de cana-de-açúcar por meio de técnicas de agrupamento em séries temporais de NDVI”. In: XV Simpósio Brasileiro de Sensoriamento Remoto, Curitiba - PR, p. 383-390.