X WORCAP
20 e 21 de outubro
Instituto Nacional de Pesquisas Espaciais - INPE
São José dos Campos - SP

Agrupamento espaço-temporal de descargas atmosféricas por método de subida de gradiente

Cesar Strauss[1], **Stephan Stephany**[2]

[1] Coordenação de Ciências Espaciais e Atmosféricas - INPE
São José dos Campos, SP - Brasil

[2]Laboratório Associado de Computação e Matemática Aplicada - INPE
São José dos Campos, SP - Brasil

cstrauss@cea.inpe.br, stephan@lac.inpe.br

Abstract. *Convective nuclei are cloud formations whose birth, evolution and dissipation are of great interest among meteorologists. These nuclei typically show a multitude of atmospheric electric discharges, whose spatio-temporal clustering correspond to entities that may be called centers of electric activity. It is possible to generate a density field of the occurrences of atmospheric electric discharges for specific time intervals and regions, being this field convenient for the visualization of the discharges. However, the evolution of this field in successive time steps do not allow to clearly distinguish the evolution of its denser regions that correspond to the centers of electric activity. Therefore, in this work, it is proposed the use of an algorithm based on gradient climb to identify centers of electric activity based on the density field of occurrences of electric discharges, enabling the visualization and the tracking of these centers.*

Resumo. *Núcleos convectivos são formações de nuvens cumulus nimbos cujo nascimento, evolução e dissipação são de grande interesse para os meteorologistas. Estes núcleos tipicamente apresentam inúmeras descargas elétricas atmosféricas, cujo agrupamento espaço-temporal corresponde a entidades chamadas centros de atividade elétrica. É possível gerar um campo de densidade de ocorrências de descargas elétricas atmosféricas para intervalo de tempo e região determinados, sendo este campo conveniente para visualização das descargas. Entretanto, a evolução deste campo em intervalos de tempo sucessivos não permite distinguir claramente a evolução de suas regiões mais densas, que correspondem aos centros de atividade elétrica. Assim, neste trabalho, propõe-se o uso de um algoritmo baseado em subida de gradiente para identificar os centros de atividade elétrica a partir do campos de densidade de ocorrências de descargas, possibilitando a visualização e rastreamento desses centros.*

Palavras-chave: *agrupamento espaço-temporal, mineração de dados, centros de atividade elétrica*

1. Introdução

Núcleos convectivos são formações de nuvens cumulus nimbus cujo nascimento, evolução e dissipação são de grande interesse por parte dos meteorologistas. O campo de densidade de descargas elétricas atmosféricas foi empregado tentativamente para rastrear esses núcleos [Politi 2005] e [Politi et al. 2006].

Nesse escopo, propomos uma metodologia para identificar e rastrear a evolução de centros de atividade elétrica usando agrupamento espaço-temporal. Este trabalho estende um algoritmo de uso geral, o algoritmo DENCLUE [Hinneburg and Gabriel 2007] de agrupamento por densidade para tratar esse tipo de dado espaço-temporal.

Os dados utilizados neste trabalho foram coletados pela rede RINDAT (Figura 1), cedidos pelo CPTEC. Para validar o método, escolheu-se intervalos de 24 horas contínuas em dois dias diferentes (16/03/2007 e 01/10/2008) com cobertura nacional. Os dados estão organizados em tabelas no padrão UALF (Universal ASCII Lightning Format), que consiste em arquivos ASCII formatados em colunas de atributos de tamanho fixo separados por espaços em branco. Para esta análise, os atributos mais relevantes são latitude, longitude e o instante de ocorrência. As coordenadas e o tempo têm resolução de fração de grau e de nanosegundo respectivamente.

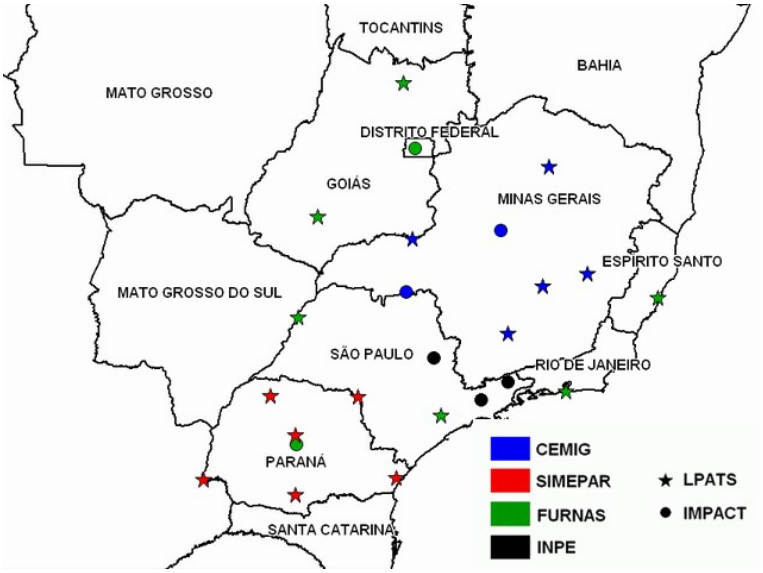

Figura 1. Localização dos sensores da rede RINDAT

2. DENCLUE

2.1. Estimador de densidade de núcleo gaussiano

A densidade de probabilidade no espaço de atributos pode ser estimada como uma função dos objetos $\boldsymbol{x}_k \in X \subset \mathbb{R}^d, d \in \mathbb{N}, t = 1, \ldots, N$. A influência dos objetos no espaço de atributos é modelada como uma simples função de núcleo, como o núcleo gaussiano:

$$K(\boldsymbol{r}) = (2\pi)^{-\frac{d}{2}} \cdot \exp\left[-\frac{\boldsymbol{r}^2}{2}\right] \tag{1}$$

onde $\boldsymbol{r}$ é o raio normalizado para uma gaussiana de desvio padrão unitário.

A soma de todos os núcleos (com a normalização apropriada) dá uma estimativa da probabilidade de qualquer ponto $\boldsymbol{x}$ no espaço de atributos:

$$\hat{p}(\boldsymbol{x}) = \frac{1}{Nh^d} \sum_{k=1}^{N} K(\boldsymbol{r}) \tag{2}$$

onde

$$\boldsymbol{r} = \frac{\boldsymbol{x} - \boldsymbol{x}_k}{h} \tag{3}$$

A estimativa $\hat{p}(\boldsymbol{x})$ tem todas as propriedades, como diferenciabilidade, da função de núcleo original. O parâmetro $h > 0$ especifica o grau de suavização de um objeto no espaço de parâmetros. Quando h é grande, um objeto estende a sua influência a regiões mais distantes. Quando h é pequeno, um objeto afeta somente a vizinhança local. [Scott 1992] e [Silverman 1986] discutem estratégias para determinar h de maneira automática.

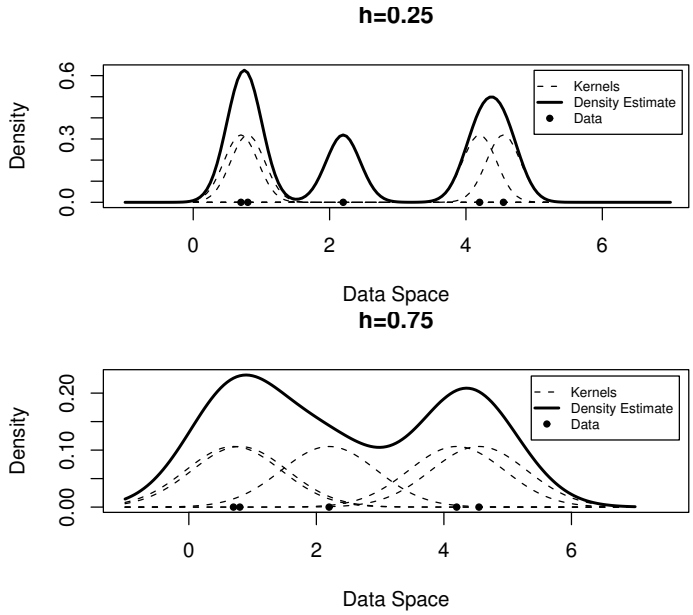

Figura 2. Estimativa de densidade de núcleo gaussiano unidimensional com valores diferentes do parâmetro de suavização h [Hinneburg and Gabriel 2007]

2.2. O método DENCLUE

O método DENCLUE [Hinneburg and Gabriel 2007] define um agrupamento pelo máximo local da função de densidade de probabilidade estimada. Uma heurística de subida de encosta é iniciada em cada objeto, que atribui o objeto ao máximo local. No caso do núcleo gaussiano, a subida de encosta é guiada pelo gradiente de $\hat{p}(\boldsymbol{x})$, que toma a forma:

$$\nabla \hat{p}(\boldsymbol{x}) = \frac{1}{h^{d+2}N} \sum_{k=1}^{N} K\left(\frac{\boldsymbol{x} - \boldsymbol{x}_k}{h}\right) \cdot (\boldsymbol{x} - \boldsymbol{x}_k) \tag{4}$$

No método DENCLUE 1.0, a subida de encosta toma a seguinte forma:

$$\boldsymbol{x}^{(l+1)} = \boldsymbol{x}^{(l)} + \delta \, \frac{\nabla \hat{p}(\boldsymbol{x}^{(l)})}{\|\nabla \hat{p}(\boldsymbol{x}^{(l)})\|_2} \tag{5}$$

onde δ é um passo fixo escolhido pelo usuário.

Um segundo método, o DENCLUE 2.0, utiliza um passo adaptativo. Fazendo o gradiente igual a zero em (4) e rearranjando, obtemos:

$$\boldsymbol{x}^{(l+1)} = \frac{\sum_{k=1}^{N} K\left(\frac{\boldsymbol{x}^{(l)} - \boldsymbol{x}_k}{h}\right) \boldsymbol{x}_k}{\sum_{k=1}^{N} K\left(\frac{\boldsymbol{x}^{(l)} - \boldsymbol{x}_k}{h}\right)} \tag{6}$$

A Figura 3 compara a eficiência dos dois métodos citados (com passo fixo e passo adaptativo). Nota-se que o método adaptativo (DENCLUE 2.0) converge mais rapidamente para o máximo local do agrupamento.

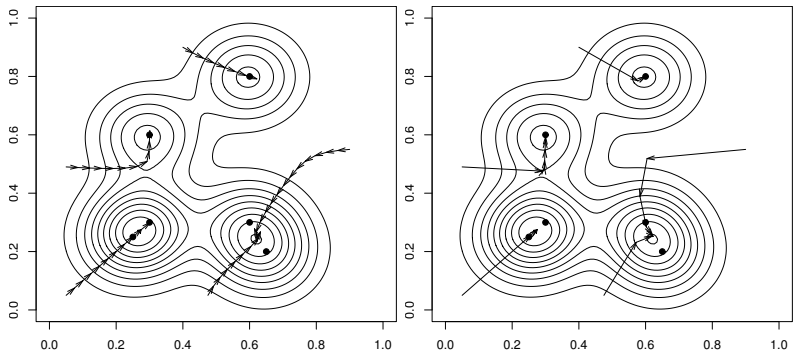

Figura 3. (esquerda) Subida de encosta por gradiente (DENCLUE 1.0), (direita) Subida de encosta com passo adaptativo (DENCLUE 2.0) [Hinneburg and Gabriel 2007]

Uma subida de encosta é iniciada a partir de cada objeto $\boldsymbol{x}_k \in X$ e itera até que se atinja um limiar específico, isto é:

$$\frac{\hat{p}(\boldsymbol{x}_k^{(l)}) - \hat{p}(\boldsymbol{x}_k^{(l-1)})}{\hat{p}(\boldsymbol{x}_k^{(l)})} \leq \epsilon \tag{7}$$

Um ponto final alcançado pela subida de encosta é denotado por $\boldsymbol{x}_k^* = \boldsymbol{x}_k^{(l)}$ e a soma dos tamanhos dos n últimos passos é $s_k = \sum_{i=1}^{n} \|\boldsymbol{x}_k^{(l-i+1)} - \boldsymbol{x}_k^{(l-i)}\|_2$. O inteiro n é um parâmetro da heurística. Para um $\epsilon > 0$ apropriado, podemos assumir que os pontos finais $\boldsymbol{x}_k^*$ estão perto do respectivo máximo local. Tipicamente, o tamanho dos passos tem uma forte tendência decrescente antes do critério de convergência ser alcançado. Assim, podemos assumir que o verdadeiro máximo está a uma distância menor que s_k do ponto $\boldsymbol{x}_k^*$. Portanto, objetos pertencentes ao mesmo agrupamento tem pontos finais $\boldsymbol{x}_k^*$ e $\boldsymbol{x}_{k'}^*$ cuja distância é menor que $s_k + s_{k'}$. Isso permite rotular os objetos de acordo com o máximo local a que pertencem.

3. Agrupamento espaço-temporal usando DENCLUE

O método DENCLUE, descrito na seção 2, pode ser estendido para dados espaço-temporais.

Seja $t_k \in T$ o instante de ocorrência do evento $\boldsymbol{x}_k$. Podemos dividir o intervalo de tempo T em intervalos de tempo $T_j = [t_0 + n\Delta t, t_0 + n\Delta t + \Delta j]$, onde t_0 é o instante inicial, Δj é a duração do intervalo e Δt é o deslocamento temporal de uma intervalo para outro. Assumimos que $\Delta j > \Delta t$, ou seja, os intervalos se sobrepõem. Então $X_j = \{\boldsymbol{x}_k | t_k \in T_j\}$ são os eventos correspondentes a cada intervalo, onde cada evento $\boldsymbol{x}_k$ pode pertencer a mais de um intervalo X_j.

Desejamos particionar os $\boldsymbol{x}_k$ em agrupamentos espaço-temporais G_r. Para cada intervalo X_j, faça:
1) Executar o algoritmo DENCLUE sobre X_j, obtendo-se os máximos locais $\boldsymbol{x}_k^*$.
2) Para cada $\boldsymbol{x}_k \in X_j$ que não pertença a nenhum agrupamento G_r, faça:
2.1) Se existe $\boldsymbol{x}_{k'} \in G_r$, onde $\|\boldsymbol{x}_k^* - \boldsymbol{x}_{k'}^*\|_2 < s_k + s_{k'}$, então faça $G_r \leftarrow G_r \cup \{\boldsymbol{x}_k\}$
2.2) Senão, crie um novo agrupamento $G_{r'} \leftarrow \{\boldsymbol{x}_k\}$

4. Metodologia

4.1. Pré-processamento

Inicialmente, fez-se a leitura do arquivo contendo os registros de dados de descargas no formato UALF, convertendo cada linha da tabela numa representação interna. Em seguida, filtrou-se os eventos conforme a área, intervalo de tempo e atributos desejados (por exemplo, somente as descargas nuvem-solo). Finalmente, segmenta-se os dados em intervalos de tempo.

4.2. Agrupamento

Aplicou-se o método DENCLUE 2, em cada intervalo de tempo, de acordo com a seção 2. O rotulamento espaço-temporal foi feito de acordo com a extensão espaço-temporal do DENCLUE 2 (seção 3). A partir daí, os centros de atividade elétrica foram determinados pelo método a seguir.

Podemos estimar, em cada intervalo de tempo X_j, a densidade de objetos devido a cada agrupamento G_r:

$$\hat{p}(\boldsymbol{x}, X_j, G_r) = \frac{1}{Nh^d} \sum_{x_k \in X_j \cap G_r} K\left(\frac{\boldsymbol{x} - \boldsymbol{x}_k}{h}\right) \tag{8}$$

Definimos um centro C_{rj}, correspondente a um agrupamento G_r e um intervalo X_j, como uma região do espaço onde a densidade é maior que um certo ξ_{rj}:

$$C_{rj} = \{\boldsymbol{x} \in X | \hat{p}(\boldsymbol{x}, X_j, G_r) > \xi_{rj}\} \tag{9}$$

onde ξ_{rj} é um parâmetro que define o tamanho dos centros obtidos.

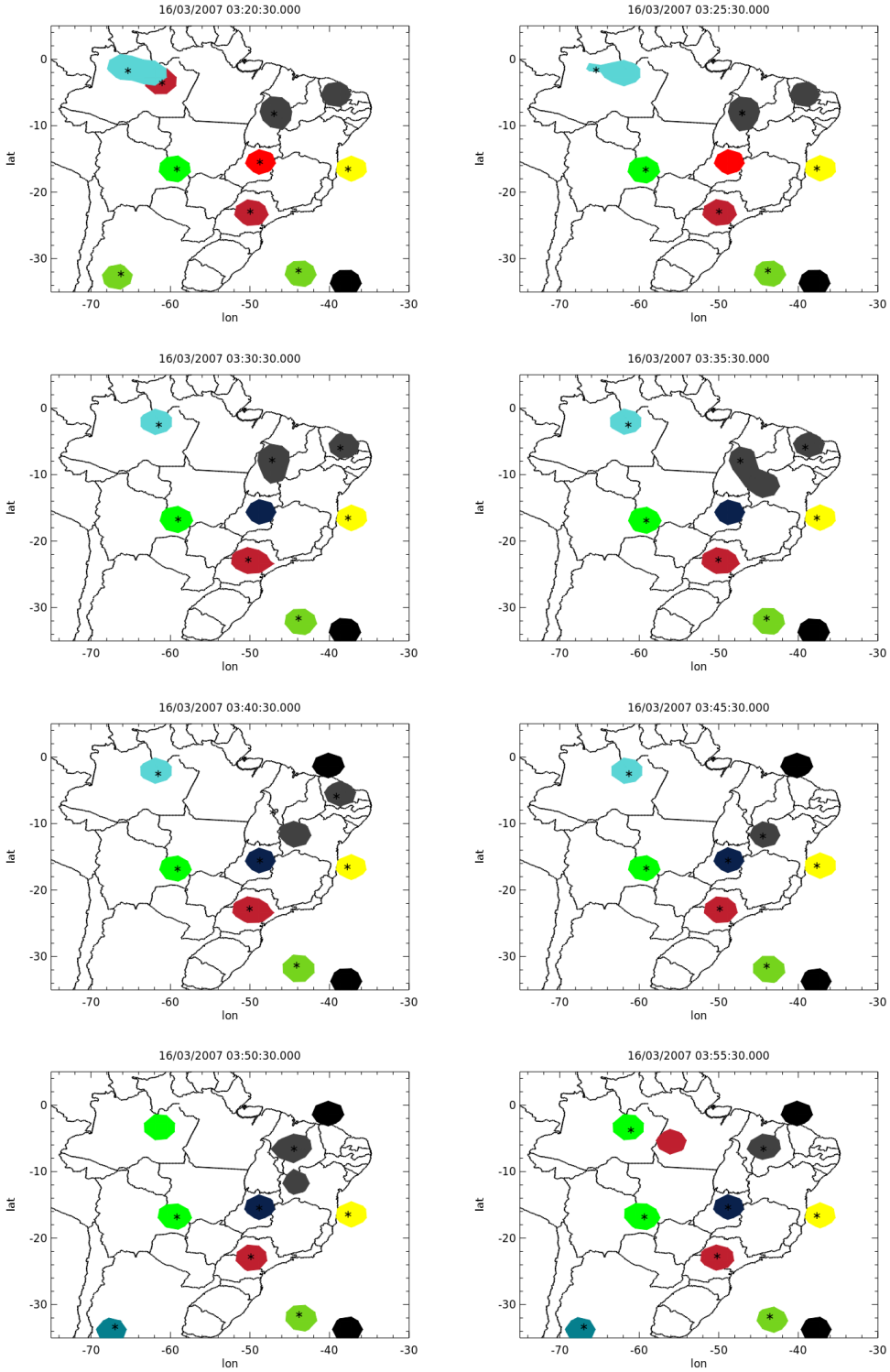

Figura 4. Resultado do método DENCLUE 2. Dados de 16/03/2007

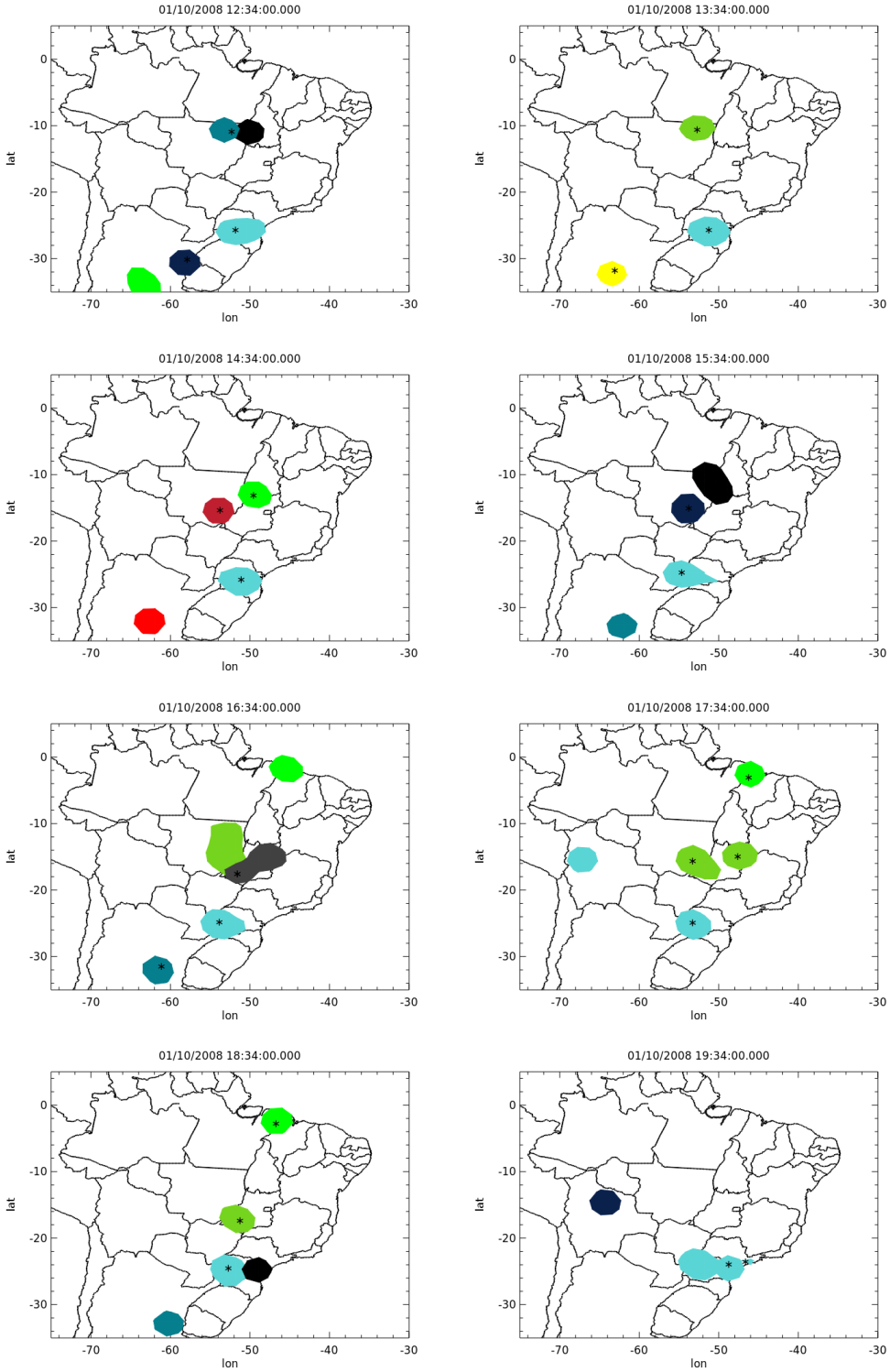

Figura 5. Resultado do método DENCLUE 2. Dados de 01/10/2008

5. Resultados

Fixou-se o intervalo de tempo em 12 minutos, com sobreposição de 6 minutos, num total de 240 intervalos. O limiar que define o tamanho do centro foi fixado em 50% do máximo de densidade do agrupamento. Para cada intervalo de tempo, gerou-se um gráfico dos centros sobrepostos a um mapa do território nacional. Os centros receberam cores aleatórias para diferenciá-los, mas o tom da cor em si não tem significado próprio. Utilizou-se um codificador de vídeo para gerar uma animação do resultado, a uma taxa de 30 quadros por segundo, onde cada quadro da animação corresponde a um intervalo de tempo. A duração total da animação é de 10 segundos, compreendendo um período contínuo de 24 horas. As figuras 4 e 5 contém alguns quadros das animações criadas.

Nota-se que a identidade de cada centro de atividade elétrica se preserva ao longo do tempo, ou seja, o mesmo centro (identificado por uma cor individual) pode ser encontrado em quadros sucessivos, aproximadamente na mesma posição. Uma possível aplicação é associar atributos ao centros, como posição, área, número de descargas, carga total, etc. Ademais, o resultado em forma de animação é adequado para a visualização dos centros.

Uma possível forma de validação desses resultados seria repetir o experimento usando outra forma de agrupamento, como k-means, e comparar os resultados através de estratégias de validação de cluster baseadas na separação e tamanho de cluster [Halkidi et al. 2001]. Outra possibilidade é tentar estabelecer a correlação entre esses agrupamentos e núcleos convectivos observados em imagens de satélites meteorológicos, como no caso do satélite GOES 12, sempre com base na hipótese de que os centros de atividade elétrica sejam representativos de atividade convectiva significativa.

6. Conclusão

Neste trabalho, estendeu-se o método de agrupamento DENCLUE, baseado no gradiente do campo de densidade, para dados espaço-temporais. Em seguida, aplicou-se o método à dados de descargas da rede RINDAT, permitindo a identificação e acompanhamento de cada centros de atividade elétrica.

Referências

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, 17:107–145.

Hinneburg, A. and Gabriel, H. (2007). Denclue 2.0: Fast clustering based on kernel density estimation. *Lecture Notes in Computer Science*, 4723:70.

Politi, J. (2005). Implementação de uma metodologia para mineração de dados aplicada ao estudo de núcleos convectivos. Master's thesis, Computação Aplicada, INPE, INPE-14165-TDI/1082.

Politi, J., Stephany, S., Domingues, M. O., and Junior, O. M. (2006). Mineração de dados meteorológicos associados a atividade convectiva empregando dados de descargas elétricas atmosféricas. *Revista Brasileira de Meteorologia*, 21(2):232–244.

Scott, D. W. (1992). *Multivariate Density Estimation - Theory, Pratice and Visualization*. John Wiliey & sons, Inc., New York.

Silverman, B. W. (1986). *Density Estimtion for Statistics and Data Analysis*. Chapman and Hall, London.