Revista Brasileira de Meteorologia, v.21, n.2, 232-244, 2006

MINERAÇÃO DE DADOS DE METEOROLÓGICOS ASSOCIADOS A ATIVIDADE CONVECTIVA EMPREGANDO DADOS DE DESCARGAS ELÉTRICAS ATMOSFÉRICAS

JACQUES POLITI[1], STEPHAN STEPHANY[2], MARGARETE O. DOMINGUES[2] e ODIM MENDES JUNIOR[3]

[1] Programa de Pós-graduação em Computação Aplicada (CAP/INPE)
jpoliti@lac.inpe.br
[2] Laboratório Associado de Computação e Matemática Aplicada (LAC/INPE)
stephan@lac.inpe.br; margarete@lac.inpe.br
[3] Divisão de Geofísica Espacial (DGE/INPE) - Instituto Nacional de Pesquisas Espaciais (INPE)
Av. dos Astronautas, 1758, Jardim da Granja, CEP 12227-010 São José dos Campos, SP – Brasil
odim@dge.inpe.br

Recebido Maio 2005 - Aceito Setembro 2005

RESUMO

Neste trabalho implementa-se a mineração de dados com um algoritmo de conjuntos aproximativos (*rough sets*) para análise quantitativa de dados meteorológicos na presença de atividade convectiva mais intensa. Foram empregados dados de ocorrências de descargas atmosféricas nuvem-solo, índices de instabilidade derivados de perfis atmosféricos obtidos de radiossondagens nos aeroportos do Centro-Sul do Brasil e dados de um modelo de previsão numérica de tempo, sendo parte desses dados obtida de campanhas do Experimento Interdisciplinar do Pantanal. Devido à grande quantidade de dados de descargas disponíveis, foi utilizada uma técnica de redução de dados por meio do agrupamento espaço-temporal das ocorrências em entidades denominadas centros de atividade elétrica (CAEs). Os CAEs foram associados à presença de atividade convectiva e usados para rastreá-la. Esta aplicação de mineração de dados buscou encontrar regras quantitativas potencialmente úteis a partir de tabelas de dados meteorológicos associados aos CAEs. A abordagem proposta de mineração de dados mostrou-se viável, sendo encontradas diversas regras quantitativas que indicaram padrões meteorologicamente coerentes, encorajando novos desenvolvimentos.

Palavras-chave: mineração de dados, conjuntos aproximativos, sistemas convectivos, descargas elétricas atmosféricas.

ABSTRACT: METEOROLOGICAL DATA MINING ASSOCIATED TO CONVECTIVE ACTIVITY EMPLOYING ELECTRIC ATMOSPHERIC DISCHARGE DATA

In this work, a data mining implementation using the rough sets theory is employed to perform a quantitative analysis of meteorological data in presence of more intense convective activity. This implementation employed data of occurrences of cloud-to-ground atmospheric electrical discharges, atmospheric profiles obtained by radiosonde in Middle-South airports of Brazil and numerical weather forecasting model data. Some of these data was obtained from campaigns of the Pantanal Interdisciplinary Experiment. Due to the high amount of discharge occurrence data, a data reduction technique was employed. It consists of the space-temporal grouping of these occurrences in entities named Electrical Activity Centers (CAEs). The CAEs were associated to the occurrence of convective activity and employed to trace it. The aim of this data mining application was to search useful quantitative rules from tables containing meteorological data associated to the CAEs. The proposed approach showed to be feasible and allowed to find several quantitative rules that can be associated to consistent meteorological patterns encouraging further enhancements.

Keywords: Data mining, rought sets, convective system, atmospheric electrical discharges

1. INTRODUÇÃO

Analisar a crescente quantidade de dados meteorológicos, gerados por sensores ou por simulações, não é uma tarefa trivial. Assim, técnicas computacionais avançadas mostram-se necessárias para descobrir correlações potencialmente úteis entre os diversos dados ou encontrar regras quantitativas associadas aos mesmos. Esse é um dos objetivos da mineração de dados. Sua aplicação ao estudo de núcleos convectivos é de interesse em pesquisa e no setor operacional em Meteorologia. Em particular, os dados de ocorrências de descargas elétricas atmosféricas nuvem-solo começam a se tornar disponíveis para uma grande extensão geográfica do Brasil com resolução temporal da ordem de dezenas de nanossegundos. Outros dados tradicionalmente utilizados no estudo de núcleos convectivos são os índices de instabilidade provenientes de perfis atmosféricos obtidos por radiossondagem, dados de modelos de previsão numérica de tempo e dados de imagens de satélites e de radares.

Em geral, os núcleos convectivos estão associados a um ou mais aglomerados de nuvens Cumulonimbus (Cb). Essas nuvens são caracterizadas pelo forte movimento vertical, grande extensão vertical, cerca de 16 km a 18 km de altura nos trópicos, e atividade elétrica intensa (MacGorman e Rust, 1998). Essas nuvens geram descargas elétricas, que são conseqüência das cargas elétricas que se acumulam devido às colisões entre diferentes tipos de partículas tais como os cristais de gelo e granizo, atingindo às vezes a carga elétrica total de até centenas de Coulombs (Uman, 1987; Volland, 1984). Essas descargas ocorrem quando o campo elétrico excede localmente a capacidade isolante do ar, acima de 400 kV/m. As descargas que atingem o solo são denominadas de descargas nuvem-solo (NS). Os relâmpagos são formados por uma ou mais dessas descargas elétricas, de caráter transiente, portando uma alta corrente elétrica, em geral, superior a várias dezenas de quiloampéres (Mendes e Domingues, 2002).

Um dos métodos atuais utilizados para o acompanhamento dos núcleos convectivos é feito por meio de imagens geradas por radares e satélites metereológicos geo-estácionários. Entretanto, a área de cobertura desses radares é pequena e não é capaz abranger toda a extensão geográfica do nosso país, prejudicando análises espaciais detalhadas de algumas regiões específicas. Por outro lado, as imagens geradas por esses satélites são coletadas em intervalos de tempo em torno de 30 minutos, sendo então, transmitidas e processadas, fazendo com que não estejam disponíveis em tempo quase real. Devido a essa freqüência de amostragem, perde-se a resolução temporal, tornando mais difícil uma análise mais detalhada de um determinado núcleo convectivo. Por outro lado, os dados de descargas elétricas do tipo nuvem-solo estão disponíveis com uma freqüência muito maior (menos de mili-segundo) do que as imagens de satélites, além de possuir uma maior região de abrangência em nosso país em relação às imagens de radar, e portanto podem ser utilizados como uma ferramenta auxiliar na detecção e acompanhamento dos núcleos convectivos.

O agrupamento espaço-temporal de ocorrências de descargas elétricas do tipo nuvem-solo foi realizado de maneira inovadora por meio de uma técnica de análise espacial e utilizado para rastreamento de núcleos convectivos (Politi 2005). Esse agrupamento resultou em entidades denominadas centros de atividade elétrica (CAEs).

Neste trabalho, o objetivo é a utilização da mineração de dados para buscar regras quantitativas aplicáveis a dados meteorológicos na presença de atividade convectiva de Cbs, rastreada por meio dos CAEs, conforme proposto anteriormente (Politi et al, 2004a, 2004b, 2004c). Os dados meteorológicos considerados incluem, além dos dados de ocorrências de descargas, índices de instabilidade derivados de perfis atmosféricos obtidos por radiossondagens nos aeroportos do Centro-Sul do Brasil e dados da reanálise do modelo do National Centers for Environmental Prediction (NCEP), tais como velocidade e direção do vento, umidade relativa e temperatura. Parte desses dados foi obtida de campanhas do Experimento Interdisciplinar do Pantanal (Domingues et al, 2004). Do ponto de vista termodinâmico, os índices de instabilidade são considerados bons previsores de tempestades originadas por núcleos convectivos. Por outro lado, as variações desses índices e sua análise quantitativa são influenciadas por fatores como estações do ano, localidade, tipo do fenômeno envolvido, etc.

Adicionalmente, este trabalho lança bases visando a caracterização de sistemas convectivos por meio de dados de descargas, o que forneceria uma informação auxiliar importante sobre a região afetada pela parte mais severa da tempestade, devido à sua estrutura termo-eletrodinâmica. Nesse sentido, outros dados seriam incorporados, de forma a enriquecer a informação associada aos CAEs, não mais restritos apenas ao rastreamento de sistemas convectivos, mas à também à sua caracterização.

A Seção 2 apresenta a metodologia e os dados utilizados. A Seção 3 apresenta os resultados alcançados e a Seção 4, as considerações finais.

2. METODOLOGIA E DADOS

Descrevem-se a seguir, na Seção 2.1, a metodologia utilizada para agrupar os dados de ocorrências de descargas em CAEs e para associá-los aos dados meteorológicos que foram utilizados. Em seguida, na Seção 2.2, é apresentada a metodologia de mineração de dados utilizada.

A mineração de dados propriamente dita foi realizada por meio de um algoritmo genético do sistema ROSETTA

(*Rough Sets Toolkit for Analysis Data*), para encontrar regras de decisão que utilizam conceitos da teoria dos conjuntos aproximativos (*rough sets*), apropriada para o tratamento da incerteza e esparsidade dos dados. Nessa aplicação, associada a núcleos convectivos, a existência de CAEs com densidade de descargas intensa foi utilizada como atributo de decisão, enquanto que os dados meteorológicos, como atributos de informação.

Os dados de ocorrências de descargas foram coletados pela Rede Integrada Nacional de Descargas Atmosféricas (RINDAT, Beneti et al., 2000) e agrupados em CAEs. Os dados de radiossondagem analisados foram coletados em fevereiro e março de 2002 nos aeroportos do Centro-Sul do Brasil. Os demais dados meteorológicos foram obtidos a partir dos dados de reanálise do modelo do NCEP (Kalnay et al, 1996) e são descritos a seguir. Tanto estes dados, como aqueles do RINDAT foram obtidos em consequência do trabalho realizado no Experimento Interdisciplinar do Pantanal (IPE – Interdisciplinary Pantanal Experiment), mais especificamente, a segunda campanha, IPE-2, que ocorreu no período de 14 a 23 de setembro de 1999 (Domingues et al, 2002, 2004), e a terceira, IPE-3, no período de 1º de fevereiro a 30 de março de 2002 (Domingues et al, 2003a, 2003b; Manzi et al., 2003). Estes dados incluem pressão atmosférica ao nível do mar e na superfície e componentes zonal e meridional da velocidade do vento a 10 m de altura. Incluem também as componentes zonal e meridional da velocidade do vento, a velocidade vertical, as umidades específica e relativa e a temperatura do ar nos níveis de pressão de 200 hPa, 500 hPa e 850 hPa.

Foram também utilizados índices de instabilidade atmosférica, calculados a partir de radiossondagens em aeroportos do Centro-Sul do Brasil. Esses índices são: CAPE (Convective Available Potencial Energy), K, TT (Totals) e SLI (Lift Index), discutidos em Doswell e Rasmussen(1994), Doswell et al. (1985), Williams e Renno (1993), Moncrieff e Green (1972), George (1960), Miller (1972) e Galway (1956).

2.1. Agrupamento espaço-temporal de descargas em CAEs

As ocorrências de descargas são breves e espalhadas temporal e espacialmente no caso de haver vários Cbs. Como então criar um campo de representação, ou uma representação espacial que evolua no tempo, de forma a monitorar esses sistemas convectivos associados às tempestades elétricas? A análise espaço-temporal das ocorrências de descargas elétricas exige uma forma de representação conveniente das mesmas, o que é agravado pela alta freqüência de ocorrência de descargas elétricas em uma tempestade, que gera um volume total de dados elevado. Na tentativa de resolver essa questão, foi desenvolvida, como abordagem inovadora, a metodologia de agrupamento espaço-temporal de dados de ocorrências de descargas em CAEs. No restante desta seção, exceto pelas referências específicas à teoria de *kernel estimator* (estimador de núcleo) usada para gerar os CAEs, a aplicação específica dessa metodologia para ocorrências de descargas foi baseada no trabalho de Politi (2005).

Diversas metodologias de análise espacial foram avaliadas para a representação espaço-temporal das descargas nuvem-solo: paintball (plotagem de eventos), agrupamento em grade, clustering (aglomerado), bem como técnicas baseadas em estimadores de densidade (Bailey e Gatrell, 1995). Após os testes com as diversas metodologias, não descritas aqui devido a estarem foram do presente escopo, foi escolhido um tipo de estimador de densidade chamado estimador de núcleo, em razão da representação obtida – um campo suave que melhor permitia a análise pretendida.

O estimador de núcleo tem ampla aplicabilidade em diversas áreas (Silverman, 1990; Grillenzoni, 2004; e Flahaut et al, 2003), devido às suas propriedades estatísticas e à flexibilidade de configuração de seus parâmetros. Nesta técnica, para o caso bidimensional, considera-se uma região genérica A que engloba n ocorrências observadas localizadas em $x_1, \ldots, x_n$ e define-se uma região circular de influência $S \subset A$ centrada numa localização de interesse x_0, que constitui um ponto de ocorrência, e delimitada por um raio de influência r, como esquematizado na Figura 1.

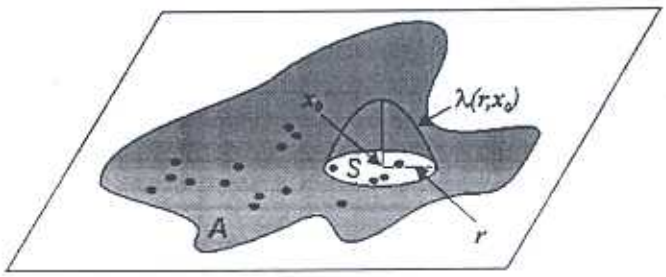

Figura 1 – Esquema ilustrativo da região de influência do estimador de núcleo.

Ajusta-se então uma função de densidade de probabilidade $\lambda(r;x0)$ sobre as ocorrências consideradas num intervalo de tempo determinado nessa região de influência S. Essa função, desconhecida, compõe uma superfície cuja altura sobre o plano bidimensional considerado será proporcional à quantidade de ocorrências por unidade de área, ponderando-as pela distância de cada ocorrência a x_0.

A função $\lambda(r;x0)$ é calculada a partir das m ocorrências localizadas em S, ajustadas por uma função de interpolação K, conhecida como estimador de núcleo da função de densidade de probabilidade $\lambda(r;x0)$, conforme a equação (1):

$$\lambda(r, x_0) = \frac{1}{mr^2} \sum_{i=1}^{m} K(y_i) \qquad (1)$$

em que $y_i = d(x_0, x_i)/r$, na qual $d(x_0, x_i)$ é a distância euclidiana de cada ponto da ocorrência x_i à localização de interesse x_0.

O raio de influência r define a vizinhança do ponto a ser interpolado, controlando a "suavidade" da superfície gerada, é também chamado *smoothing parameter*. Quanto maior for esse raio, mais suavizada será a superfície gerada, e vice-versa, sendo sua escolha um fator importante, pois define o diâmetro médio dos campos gerados. A função de interpolação K é também uma função de densidade de probabilidade, sendo, no entanto, conhecida e escolhida convenientemente. Segundo Epanechnikov (1969), a escolha da função de interpolação K não é crítica para o desempenho estatístico do método, mas certamente tem influência na representação obtida. Silverman (1990), abordou critérios para ajuste automático ótimo do parâmetro de suavização r, sendo que o método mais amplamente utilizado baseia-se no erro quadrático médio integrado (MISE – *mean integrated square error*). Quando $\lambda(r,x0)$ e a função K são gaussianas, pode-se demonstrar que o MISE é minimizado para $r = 1{,}06\sigma m^{-1/5}$ onde σ é o desvio padrão da amostra de dados. Essa técnica para estimação do raio de influência é freqüentemente utilizada, sendo conhecida como regra prática de Silverman (Silverman's rule of thumb, (Lee 2003)). Os testes realizados demonstraram a conveniência da escolha da função gaussiana para K, dada a suavidade dos campos obtidos, e do uso dessa regra prática, dada a delimitação mais precisa das regiões de atividade elétrica.

Primeiramente, deve-se realizar a integração temporal dos dados separando as ocorrências de descargas em intervalos de tempo compatíveis com a escala de tempo do fenômeno observado, ou seja, o ciclo das estruturas convectivas. Definida uma grade temporal, optou-se por considerar intervalos de tempo centrados nos pontos dessa grade, de forma a realizar a integração das descargas temporalmente próximas e não apenas as descargas já ocorridas. Isso traz como vantagem a possibilidade de integração com outros dados (temperatura, pressão, etc) que geralmente possuem resolução temporal mais baixa.

Em seguida, é necessária a análise da evolução espaço-temporal dos CAEs por meio de sua identificação, uma vez que um CAE pode se deslocar ou deixar de existir e, em ambos os casos, novos CAEs podem surgir. Para tal, definiu-se uma distância de máximo deslocamento do CAE e também um número máximo de intervalos de tempo em que um determinado CAE não é detectado. Isto porque pode ter ocorrido um descarregamento elétrico do sistema convectivo associado a um determinado CAE num intervalo e este somente voltar a ter atividade elétrica após alguns intervalos. Essa evolução espaço-temporal CAEs pode ser associada à própria evolução física da atividade convectiva mais intensa.

Essa metodologia foi avaliada utilizando-se imagens GOES e foi possível observar que os CAEs encontram-se dentro das regiões delimitadas pelas nuvens convectivas, e indicam quais dessas possuem atividade elétrica. Obviamente, não há uma correspondência exata entre as regiões de Cbs das imagens GOES com os CAEs, uma vez que estes correspondem a intervalos de tempo, enquanto aquelas são instantâneas. Além disto, houve a diferença de escalas e projeções cartográficas entre essas imagens e os CAEs. É conveniente ressaltar que as técnicas empregadas nesta avaliação "enxergam" cenários físicos distintos associados à atividade convectiva (imagens meteorológicas e campos de descargas elétricas). Assim, constatou-se que a metodologia proposta para obtenção dos CAEs é robusta o suficiente para o rastreamento de atividade convectiva, desde que sejam estabelecidos limiares convenientes para eliminar, por exemplo, ocorrências de descargas esparsas espacial ou temporalmente.

Assim, foi possível gerar CAEs com aparência suavizada e que com continuidade temporal de forma a se poder rastrear, mesmo de forma aproximada, o centro de atividade convectiva associada. Na Figura 2, pode-se observar um exemplo dos CAEs gerados, onde os pontos pretos representam as ocorrências de descargas elétricas e a escala de cores foi baseada na densidade de descargas, variando do vermelho (maior densidade) para o azul (menor densidade).

Figura 2 – Exemplo de CAEs encontrados na região Sudeste em 2002.

Os CAEs foram supostos como representativos de atividade convectiva e, neste trabalho, foram associados a um ou outro dos seguintes tipos de dados:

(i) Campos meteorológicos obtidos dos dados de reanálise do modelo global do NCEP em alguns níveis padrão.

(ii) Índices de instabilidade derivados de perfis atmosféricos obtidos por radiossondagem.

No caso dos índices de instabilidade, as tabelas que os associam aos CAEs, foram geradas por uma das estratégias expostas a seguir:

- **Associação das estações de radiossondagem a CAEs.** Para um dado CAE, verifica-se se uma dada estação está localizada dentro de sua área e, em caso afirmativo, associam-se todas as informações desta estação a este. Os CAEs são então agrupados em uma tabela contendo um resumo de suas características, tais como a posição espacial de seu centro, área, densidade, número de ocorrências de descargas e índices de instabilidade derivados dos dados da estação de radiossondagem associada.
- **Associação de CAEs a estações de radiossondagem.** Para uma dada estação, dentro de um raio definido (100 ou 200 km), define-se um atributo de decisão relativo à existência de CAEs dentro desse raio. As estações são agrupadas numa tabela que contém todas as informações das mesmas, além desse atributo de decisão.

As tabelas geradas para ambos os tipos de dados (NCEP e índices de instabilidade) são então utilizadas como entrada para o sistema de mineração de dados ROSETTA, conforme exposto na seção seguinte.

2.2. Metodologia de mineração de dados

O processo de mineração de dados, também conhecido por KDD (*Knowledge Discovery in Databases* – Descoberta de Conhecimento em Banco de Dados) consiste basicamente de seis etapas e cada uma delas pode se sobrepor às demais (Piatetsky-Shapiro, 1991). A Figura 3 ilustra todo o processo, sendo possíveis realimentações de dados entre as etapas.

As etapas que consomem mais tempo computacional são o pré-processamento e a transformação de dados. Essas etapas visam construir uma base de dados integrada, tão confiável quanto possível, e principalmente reduzida, proporcionando um melhor desempenho ao algoritmo da etapa específica de mineração de dados.

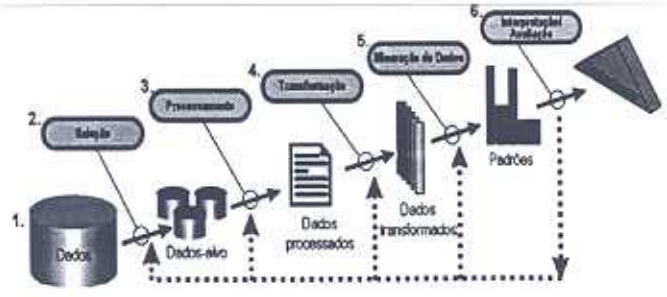

Figura 3 – Etapas do ciclo de mineração de dados. *Fonte: Fayyad (1996).*

A metodologia proposta pode ser dividida em duas partes distintas. A primeira diz respeito às etapas de pré-processamento e transformação dos dados de ocorrências de descargas elétricas em CAEs e sua associação com os dados meteorológicos, que foi implementada inicialmente no ambiente MATLAB®, enquanto que a segunda refere-se à mineração de dados propriamente dita utilizando uma ferramenta aberta, o sistema ROSETTA (*Rough Sets Toolkit for Analysis Data*, ver Øhrn, 1999), que utiliza a teoria dos conjuntos aproximativos (*Rough Sets*).

A teoria dos conjuntos aproximativos foi desenvolvida por Zdzislaw Pawlak no começo da década de 80 para lidar com dados incertos e vagos em aplicações de Inteligência Artificial (Pawlak, 1982). Essa teoria tem se mostrado como uma base teórica para a solução de muitos problemas com mineração de dados, principalmente no que diz respeito à redução de dados.

Uma das vantagens desta teoria é que não necessita de nenhuma informação preliminar ou adicional sobre os dados, ao contrário do que acontece na teoria dos conjuntos nebulosos que necessita de uma função de pertinência para transformar os dados reais em valores nebulosos (Chen, 2001). Além dessa característica, podem-se citar também a obtenção de conjuntos mínimos de dados que possibilitam a geração de regras de decisão, o tratamento quantitativo da incerteza e métricas estatísticas para avaliar a importância das regras.

A teoria dos conjuntos aproximativos baseia-se principalmente nas relações de indiscernibilidade ou similaridade entre os objetos (registros). Essas relações permitem que um sistema de informação (registros + atributos condicionais) seja particionado em classes de equivalência, de acordo com determinados subconjuntos de atributos. Ao expandir o conceito de sistema de informação para sistema de decisão (registros + atributos condicionais + atributos de decisão), podem-se obter situações ou regras não-determinísticas, como por exemplo, registros que contenham os mesmos valores de atributos condicionais, mas com valores de atributos de decisão diferentes (inconsistências). Devido à necessidade de quantificar esse não determinismo, surgem os conceitos de aproximação inferior e aproximação superior. Na primeira, os elementos do conjunto certamente pertencem à determinada classe e na segunda os elementos possivelmente pertencem à classe. A diferença entre aproximação superior e aproximação inferior forma a região conhecida como borda ou fronteira, o que auxilia a visualizar tendências possíveis.

O sistema ROSETTA é um conjunto de componentes de software escrito na linguagem C++ utilizados para análise de dados. O sistema foi desenvolvido em um esforço cooperativo entre o grupo *Knowledge Discovery Group* da NTNU (*Norwegian University of Science and Technology*), na Noruega e o *Logic Group* da Universidade de Varsóvia, Polônia. O ROSETTA é capaz de suportar todo o ciclo de mineração de

dados, incluindo o pré-processamento e transformação dos dados.

Os dados contidos nas tabelas resultantes da associação dos CAEs com os dados meteorológicos são agrupados em classes discretas. Essas tabelas constituem a base de dados utilizadas pelo algoritmo de mineração de dados propriamente dito, que efetua as reduções de atributos por meio de conjuntos aproximativos. Geram-se então regras do tipo "*if-then*", que estão sempre associadas a medidas quantitativas que auxiliam na determinação da importância de cada regra com base em sua cobertura estatística, como exemplificado na Tabela 1. Essas regras representam faixas de variação de um ou mais parâmetros meteorológicos (variáveis meterológicas ou índices de instabilidade) na presença de atividade convectiva, ou seja na presença de CAEs com densidade de descargas intensa.

Na primeira coluna da Tabela 1, tem-se exemplos de regra gerada pelo algoritmo genético de redução do sistema ROSETTA no caso de dados de estações de radiossondagem (índices de instabilidade). Cada regra é separada em duas partes pelo símbolo "=>". O lado esquerdo da regra (LHS, *Left Hand Side*), apresenta os atributos condicionais ligados por "*and*" ("E" lógico), e o lado direito (RHS, *Right Hand Side*), apresenta o atributo de decisão, no caso a existência de atividade convectiva (CAEs com densidade de descargas intensa). Os números entre parênteses/colchetes correspondem aos valores dos atributos e o "*" significa infinito.

Como usual, parênteses delimitam intervalos abertos, e os colchetes, intervalos fechados. Na coluna "suporte" **(SUP)**, tem-se o número de registros da base que satisfazem o RHS da regra, complementado pelo correspondente percentual do total de registros na coluna "cobertura" **(COB)**. Na primeira linha da tabela, por exemplo, pode-se interpretar a regra como: "Em 71% dos casos em que houve atividade convectiva próxima a estações de radiosondagem, o índice K foi MAIOR que 32 ".

Tabela 1 – Exemplos de regras do tipo IF-THEN geradas

REGRA	SUP	COB (%)
K([32, *)) => atividade convectiva	68	71
SLI([*, -1)) => atividade convectiva	54	56
K([32,*))AND T([44,*)) => atividade convectiva	52	54

Note-se que a atividade convectiva é detetada/rastreada pelos CAEs obtidos dos dados de ocorrências de descargas e, aqui, supõe-se existência de atividade convectiva (eletricamente ativa) quando os CAEs tem densidade de descargas acima de um certo limiar. Entretanto, o RHS de cada regra considera todos os CAEs, mas é preciso levar em conta apenas aqueles com densidade acima do referido limiar. Assim, para cada regra gerada pela mineração de dados, pode-se usar uma outra grandeza, a "cobertura RHS", que consiste na razão entre os casos que verificam o LHS e aqueles em que o atributo de decisão é verificado, i.e. CAEs com densidade de descargas suficientemente alta.

3. RESULTADOS

A seguir, apresentam-se os resultados referentes à aplicação da mineração de dados para a busca de regras quantitativas para parâmetros meteorológicos (variáveis meterológicas ou índices de instabilidade), associadas à presença de atividade convectiva, conforme a metodologia descrita anteriormente. Os CAEs com densidade de descargas intensa indicam a presença de atividade convectiva, sendo utilizados como atributo de decisão, enquanto os parâmetros meteorológicos, como atributos de informação. A mineração foi realizada separadamente para os dados do NCEP e para os dados de radiossondagem.

3.1. Casos de teste efetuados

De forma a buscar regras quantitativas para os parâmetros meteorológicos, na presença de atividade convectiva, foram realizados 16 testes. O número de parâmetros de controle de teste e de parâmetros meteorológicos é elevado, permitindo com isso uma grande quantidade de combinações possíveis. Sem informações adicionais sobre os parâmetros de controle e do problema físico propriamente dito, a escolha desses parâmetros, bem como a seleção dos parâmetros meteorológicos, torna-se muito complexa. Alguns parâmetros foram considerados empiricamente, enquanto que outros, graças ao auxílio de especialistas, com base em pesquisas bem reconhecidas (e.g. MacGorman e Rust, 1998; Cotton e Anthes, 1989).

No tocante à obtenção dos CAEs, o parâmetro raio de influência, na abordagem do estimador de núcleo, foi definido automaticamente pela regra prática de Silverman. Um filtro para as descargas esparsas foi definido empiricamente com o valor de 10%, baseado em comparações visuais entre os testes realizados. Foi especificado um limiar de densidade de descargas, acima do qual supõe-se a existência de atividade convectiva. Todos os testes utilizaram como tamanho de célula da grade bidimensional o valor definido de 0,3° (cerca de 33 km), no intuito de obter uma representação mais precisa dos CAEs. Os testes foram realizados na região delimitada pelas latitudes mínima e máxima de -30° à -10° respectivamente, e pelas longitudes mínima e máxima de -60° à -35°, como pode ser observado na Figura 4.

Figura 4 – Região de análise para os testes de mineração de dados efetuados.

A Tabela 2 descreve os testes efetuados de mineração de dados, efetuados com dados de radiossondagem ou do NCEP. Em ambos os casos, para tornar viável a análise proposta, os parâmetros meteorológicos devem ser discretizados, sendo sua discretização: binária (+/–) ou então definindo-se um número de intervalos (2 ou 3) correspondentes a faixas de valores "baixo/alto" ou "baixo/médio/alto", respectivamente.

Os parâmetros de controle dos testes apresentados na Tabela 2 são descritos a seguir:

P1) Campanha IPE: Nos testes numerados de 1 à 4, que utilizam dados de reanálise do modelo NCEP, o período corresponde ao IPE-2 enquanto que, nos testes numerados de 5 a 8, ao IPE-3.

P2) Dados: "E" indica que os CAEs foram associados a dados do modelo NCEP e "R" indica que foram associados aos índices de instabilidade obtidos das estações de radiossondagem.

P3) Timestep: O tempo de integração, expresso em horas, foi escolhido com o auxílio de um meteorologista, no intuito de não ultrapassar a duração típica de um núcleo convectivo.

P4) Área de influência: indica a distância máxima em graus para verificação de pertinência entre CAEs e estações de radiossondagem.

P5) Deslocamento: Consiste em deslocar a faixa de integração dos dados de descargas elétricas em CAEs, de um determinado número de horas. No caso da associação de CAEs com dados do NCEP não houve necessidade de deslocamento, pois a resolução temporal desse modelo é de 6 horas, possibilitando a análise de todos os períodos de um dia. Por outro lado, os dados das estações de radiossondagem são obtidos a cada 12 ou 24 horas, às 00 UTC ou 12 UTC. Assim, eventualmente, os dados das estações de radiossondagem podem compreender períodos em que a atividade elétrica é baixa, devido a essa resolução temporal. Por esse motivo, em alguns testes, a integração dos CAEs foi feita com antecedência de 6 horas em relação aos horários de observação dos dados das estações de radiossondagem.

Os parâmetros meteorológicos utilizados, cuja discretização é mostrada na Tabela 2 são, no caso de dados do NCEP (testes 1 a 8): temperatura do ar (*temp*), umidade específica (*umes*), umidade relativa (*umrl*), tendência da pressão atmosférica ao nível do mar/superfície (*dif_psnm* e *dif_pslc*), componente zonal da velocidade do vento a 10m de altura

Tabela 2 – Descrição dos testes efetuados.

	Identificador do Teste															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Pré-processamento (CAEs) – parâmetros de controle de teste																
P1) Campanha IPE	2	2	2	2	3	3	3	3	-	-	-	-	-	-	-	-
P2) Dados	E	E	E	E	E	E	E	E	R	R	R	R	R	R	R	R
P3) Timestep (h)	1	3	1	3	1	3	1	3	1	1	3	3	1	1	3	3
P4) Área influência (°)	-	-	-	-	-	-	-	1	1	1	1	1	1	1	1	1
P5) Deslocamento (h)	0	0	0	0	0	0	0	0	0	6	0	6	0	6	0	6
Mineração de dados - Níveis de discretização dos parâmetros meteorológicos																
Dif_psnm/pslc (hPa)	-/+	-/+	-/+	-/+	-/+	-/+	-/+	-/+								
U10m,v10m (m/s)	-/+	-/+	-/+	-/+	-/+	-/+	-/+	-/+								
uvel, vvel (m/s)	-/+	-/+	-/+	-/+	-/+	-/+	-/+	-/+								
omega (cm/s)	-/+	-/+	-/+	-/+	-/+	-/+	-/+	-/+								
demais parâmetros	2	2	3	3	2	2	3	3	2	2	2	2	3	3	3	3

(*u10m*), componente meridional da velocidade do vento a 10m de altura (*v10m*), componente zonal da velocidade do vento no nível de pressão especificado (*uvel*), componente meridional da velocidade do vento no nível de pressão especificado (*vvel*) e velocidade vertical (*omega*) e altura geopotencial (*zgeo*). Os níveis de pressão considerados foram 200 hPa, 500 hPa e 850 hPa. Na mesma tabela, a discretização adotada para os índices de instabilidade (testes 9 a 16), aparece na última linha. A umidade específica e a relativa estão associadas, mas a primeira pode ser usada como um filtro para reduzir efeitos da variação diurna da segunda. Isso se torna interessante quando se abrange uma variação temporal mais estendida, como é no caso dos testes efetuados.

No caso do dado *dif_psnm/pslc*, que indica a variação da pressão atmosférica em relação ao horário anterior de amostragem dos dados do NCEP, optou-se por uma discretização binária para se ter uma abordagem qualitativa, no caso, verificar se a pressão aumentou ou diminuiu. Analogamente, no caso das componentes zonais e meridionais da velocidade do vento e da velocidade vertical, a discretização binária indica apenas o sentido. Os demais parâmetros do NCEP, bem como os índices de instabilidade foram discretizados automaticamente em 2 ou 3 intervalos com igual número de elementos em cada um, por meio do algoritmo *Equal Frequency Binning* do software ROSETTA.

3.2. Redução de dados de descargas elétricas em CAEs

De maneira geral, o desempenho da etapa de mineração de dados propriamente dita depende muito da redução do volume de dados. A redução do número de descargas elétricas obtida pela geração das entidades denominadas CAEs foi efetiva e possibilitou a integração com os outros dados meteorológicos analisados, tornando viável a mineração de dados por meio do sistema ROSETTA.

Esta redução pode ser observada na Figura 5, na qual aparecem gráficos que representam os valores absolutos do número de descargas NS e o número de CAEs correspondentes para os testes descritos na Tabela 2. Foi mostrada nessa figura a redução obtida em 8 dos casos de teste: testes 1, 2, 5, 6, 9, 10, 11, e 12, enumerados, respectivamente, de (a) até (h). Para cada teste, a figura apresenta o total de ocorrências de descargas, o total de descargas negativas e o de positivas, sendo apresentado, à direita de cada total, o número de CAEs correspondentes. Os demais testes não foram apresentados, pois representam os mesmos dados de descargas e correspondentes CAEs, variando-se apenas o número de intervalos de discretização utilizado no sistema ROSETTA para os dados e índices meteorológicos.

Distinguem-se CAEs referentes a ocorrências de descargas negativas, positivas e a soma de ambas. Note-se que a soma das ocorrências das descargas negativas e positivas iguala o valor total, mas o mesmo não se aplica aos CAEs.

Analisando esses resultados no tocante à redução de dados (descargas agrupadas na forma de CAEs), o sistema comportou-se de maneira eficaz reduzindo os dados iniciais em cerca de 99% .

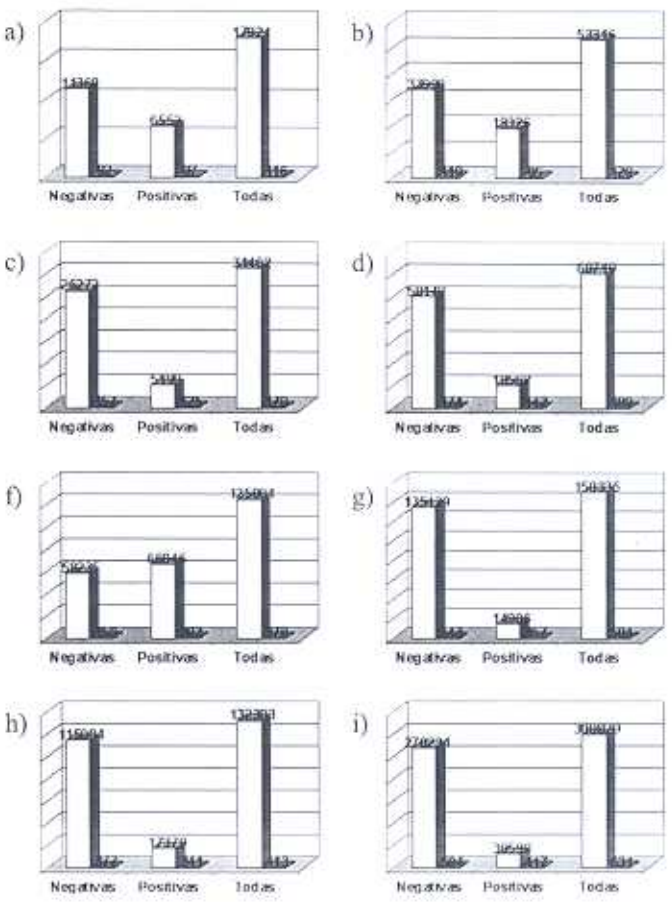

Figura 5 – Resultados da redução dos dados de descargas elétricas em CAEs. As letras indicam os respectivos testes referenciados na Tabela 2: **(a)** Teste 1; **(b)** Teste 2. **(c)** Teste 5; **(d)** Teste 6; **(e)** Teste 9; (f) Teste 10; **(g)** Teste 11; **(h)** Teste 12.

3.3. Regras derivadas pela mineração de dados

Para cada um dos 16 testes realizados, utilizaram-se 3 tabelas, para CAEs referentes a descargas negativas, positivas e ao total de ambas. Essa divisão tornou-se necessária para agregar a variável física "carga elétrica" às ocorrências de descargas negativas e positivas, expressas pelos CAEs correspondentes. Na tabela que contém CAES referentes a todas as descargas aparece apenas a densidade de ocorrências, que é função da distribuição espacial das mesmas, sem agregar/ponderar essas

ocorrências pelo valor da correspondente carga. Os domínios de valores para a carga e a densidade não podem ser definidos a priori, uma vez que apresentam variações grandes em função do número de descargas elétricas envolvidas, do intervalo do tempo de integração definido, além de outros parâmetros.

No caso dos 8 testes que utilizaram dados de radiossonagem, acrescentou-se uma quarta tabela, referente à ocorrência de CAEs nas proximidades das estações de radiosondagem. Assim, foram obtidas 32 tabelas para os últimos 8 testes e outras 24 para os demais. A mineração de dados aplicada a esse grande número de tabelas não é trivial, uma vez que é gerado um grande número de regras de decisão, o que dificulta a interpretação e avaliação dos padrões encontrados. Deve-se lembrar que para todos os conjuntos de regras, podem haver exceções que, no entanto, não comprometem o resultado e a interpretação final. Uma análise minuciosa de regras isoladas depende do nível de detalhamento que o meteorologista deseja.

Neste trabalho, foram empregadas apenas as tabelas relativas a CAEs que agregam todas as descargas, sem discriminá-las em positivas ou negativas, e portanto aparece apenas a densidade de ocorrência das mesmas.

Dentre todas as regras obtidas em cada teste, selecionaram-se apenas as 10 regras mais importantes, ou seja, as que possuíam uma maior cobertura RHS, a qual, como definida anteriormente, refere-se apenas a CAEs com densidade de descargas suficientemente alta para caracterizar a existência de atividade convectiva.

Para um melhor entendimento das regras, deve-se conhecer previamente os parâmetros meteorológicos e a forma com que foram processados/discretizados. Alguns parâmetros foram discretizados em 2 intervalos, isso indica que regras que apresentam apenas esse parâmetro e tem cobertura RHS próxima a 50% devam ser desconsideradas, pois seu significado estatístico é baixo (ou seja, seriam regras que se verificam tanto para ausência como para presença de atividade convectiva). Por outro lado, se valores próximos de 50% forem encontrados em regras que possuam mais de um parâmetro no LHS ou nas quais o parâmetro seja discretizado em mais que 2 intervalos, essa regra deverá ser considerada, pois nesses casos a cobertura RHS tende a apresentar valores mais reduzidos. A análise das regras deve ser feita de forma global, uma vez que muitos parâmetros constituintes das principais regras encontradas são comuns a vários testes.

Na Tabela 3, foi feita uma verificação de quais parâmetros são mais importantes para a tarefa de caracterização dos núcleos convectivos, ou seja, os parâmetros que mais ocorreram no conjunto total de regras. Essa tabela foi utilizada como ponto de partida para a determinação das regras mais importantes. No caso dos índices de instabilidade foi necessária a análise conjunta de todos esses índices, uma vez que a frequência de ocorrência dos mesmos nas regras é muito similar. No caso dos testes que utilizaram dados do NCEP, as variáveis mais importantes são: *dif_psnm, omeg_200, omeg_500, zgeo_500, uvel_500, zgeo_200, uvel_200, vvel_500* e *vvel_850*.

O parâmetro que teve maior destaque foi a variação de pressão (*dif_psnm/pslc*), que apresentou queda em relação às 6 horas anteriores em 100% dos casos em que houve atividade elétrica mais intensa. Além disso, as regras que contém esse parâmetro apresentaram os maiores índices de cobertura RHS, atingindo valores de até 83%. Essa queda de pressão que antecede a atividade convectiva é de conhecimento comum dos meteorologistas, mas essa associação serve para validar a metodologia proposta.

Tabela 3 – Frequência de ocorrência dos parâmetros meteorológicos no total das regras geradas para dados NCEP e dados de radiosondagem.

Número total de Regras = 461							
Regras NCEP = 240						Regras Radiossonda = 221	
parâmetro	nº de regras	parâmetro	nº de regras	parâmetro	nº de regras	parâmetro	nº de regras
dif_psnm	39	zgeo_200	22			K	148
dif_pslc	11	zgeo_500	27	Umes_200	7	TT	152
uvel_200	21	zgeo_850	19	Umes_500	8	SLI	137
uvel_500	26	temp_200	10	Umes_850	16	CAPE	124
uvel_850	7	temp_500	11	Umrl_200	1		
vvel_200	12	temp_850	14	Umrl_500	2		
vvel_500	20	omeg_200	34	Umrl_850	3		
vvel_850	20	omeg_500	34				
u10m	8	omeg_850	15				
v10m	7						

O parâmetro *omega* também foi de grande destaque na análise, apresentando valores negativos em 100% dos casos, estando também relacionado com presença de atividade convectiva, para os níveis de 200 hPa, 500 hPa e 850 hPa. Esta é outra associação de conhecimento comum dos meteorologistas, mas é interessante notar que o valor de 100% constitui uma característica marcante, dada a variabilidade espacial e temporal das tempestades e dos dados do modelo. Outra característica interessante nesse parâmetro, é que em 100% das regras, este parâmetro está associado a alguma outro, nunca aparecendo isoladamente. Isso faria com que sua cobertura RHS tendesse a apresentar valores mais reduzidos. Entretanto, foi possível observar regras com esse parâmetro em que a cobertura RHS atingiu 76%.

No caso de presença de atividade convectiva, os ventos meridionais (vvel) em 850 hPa, ou próximos da superfície (v10m), apresentam sentido Norte-Sul em 100% dos casos. Sabe-se que em muitas situações de instabilidade nas regiões Sul/Sudeste, o vento nos baixos níveis é de noroeste e que, portanto, seria interessante trabalhar com o vetor velocidade do vento, em vez das componentes zonal e meridional. Entretanto, trabalhar com esse vetor implicaria numa estrutura de correlação mais complexa, tendo-se optado neste primeiro estudo por uma abordagem mais simples, com base nas componentes zonal e meridional, discretizadas em valores positivos e negativos. Para a média troposfera (500 hPa), os ventos não são claramente definidos, apresentando 50% para o sentido Sul-Norte e 50% para o sentido Norte-Sul. E, em 200 hPa, o sentido é invertido em 83% dos casos, apresentando sentido Sul-Norte. Para os ventos zonais (uvel) nos níveis de 200 hPa e 500 hPa, 100% dos casos associados a atividade elétrica mais intensa apresentam o sentido Oeste-Leste. Essa variabilidade está associada à circulação atmosférica presente, na qual o sistema frontal possui uma ondulação, sendo a análise feita de uma maneira global e não local à região do cavado. Para o nível de 850 hPa, 57% apresentam sentido Oeste-Leste e 43% apresentam sentido Leste-Oeste. A análise meteorológica dos campos de vento desse período confirmou esses padrões de ventos.

Em seguida, apresentam-se as regras obtidas nos testes com índices de instabilidade obtidos a partir dos dados das estações de radiossondagem: CAPE, K, TT e SLI.

No caso do índice CAPE, fixou-se o valor 1000 m^2/s^2 para delimitar os valores altos dos baixos, na discretização em 2 intervalos. Em virtude dos intervalos de discretização dos outros índices de instabilidade (K, TT e SLI) serem calculados automaticamente, os valores considerados baixos e altos apresentam variações. Na Tabela 4, são apresentados os valores médios que delimitam os intervalos para cada índice, de forma a se obter 2 ou 3 intervalos de discretização.

Tabela 4 – Limites utilizados na discretização dos valores dos índices de instabilidade.

	2 intervalos	3 intervalos	
		Baixo	Alto
TT	45	44	47
K	34	32	36
SLI	-1,2	-2,2	-0,3

Observa-se que valores de corte que são considerados baixos no caso da discretização em 2 intervalos podem ser considerados altos no caso de 3 intervalos. Essa característica poderia resultar em conclusões equivocadas, ao serem feitas análises globais. Como neste caso a variação dos intervalos é relativamente pequena, é possível utilizar apenas a discretização em 2 intervalos.

Conforme mencionado anteriormente, dois tipos de tabelas foram analisadas para os dados integrados. No primeiro tipo de tabela, é verificada se a estação de radiossondagem está abrangida por um ou mais CAEs, enquanto que no segundo tipo, é verificado se há algum CAE próximo à estação, dentro de um raio definido.

Os padrões encontrados para o primeiro tipo de tabela, indicam que a ocorrência de atividade convectiva está relacionada a valores altos dos parâmetros K e TT, e a valores baixos do parâmetro SLI. Por outro lado, a não ocorrência está associada a valores opostos desses parâmetros. Ao analisar-se o segundo tipo de tabela, verifica-se que além desses padrões serem comuns, o parâmetro CAPE torna-se mais importante: nas regras, valores altos do parâmetro CAPE apareceram relacionados com cobertura RHS alta (presença de atividade convectiva) e vice-versa.

4. CONSIDERAÇÕES FINAIS

Neste trabalho, desenvolveu-se uma metodologia para mineração de dados meteorológicos associados à presença de atividade convectiva, objetivando a análise quantitativa dos dados. Foram empregados dados de ocorrências de descargas elétricas atmosféricas nuvem-solo, índices de instabilidade derivados de perfis atmosféricos obtidos por radiosondagem nos aeroportos do Centro-Sul do Brasil e dados de campos provenientes da reanálise do NCEP, sendo parte desses dados obtida de campanhas do Experimento Interdisciplinar do Pantanal.

A implementação possui duas etapas distintas: a primeira executa o pré-processamento, redução e transformação, integração e visualização dos diversos dados, enquanto que a segunda efetua a mineração de dados propriamente dita a partir de tabelas geradas na etapa anterior. Essa primeira etapa demandou a maior parte do tempo de desenvolvimento, e consume cerca de 80% do tempo de processamento.

Devido à grande quantidade de dados de ocorrências de descargas disponíveis, foi proposta uma abordagem inovadora, por meio de uma técnica de redução de dados para o agrupamento espaço-temporal das ocorrências em entidades denominadas centros de atividade elétrica (CAEs). Os CAEs com densidade de descargas acima de um certo limiar foram associados à presença de atividade convectiva eletricamente ativa e usados para rastreá-la. A utilização dessa técnica permitiu uma redução significativa do volume total de dados de descargas, que atingiu cerca de 99%. Essa redução viabilizou a integração dos dados de descargas aos demais dados meteorológicos e a subsequente mineração dos mesmos.

A etapa de mineração de dados propriamente dita, objetivou a análise quantitativa de dados meteorológicos na presença de atividade convectiva. Neste trabalho, constatou-se a adequação do sistema ROSETTA, empregado na mineração de dados, que se baseia na teoria de conjuntos aproximativos (*rough sets*). Este sistema possui algoritmos de discretização para redução do número de atributos e permite geração eficiente de regras de decisão, além de simplificar a avaliação da importância dos padrões encontrados. Estes apresentam-se na forma de regras "*if-then*" que associam quantitativamente dados meteorológicos à ocorrência de atividade convectiva, a qual é caracterizada por CAEs com densidade de descargas suficientemente alta. Estas regras correlacionam, portanto dados de descargas elétricas com dados meteorológicos com ênfase na presença de atividade convectiva.

A metodologia proposta de mineração de dados possibilitou encontrar de forma automática e quantitativa alguns padrões de conhecimento geral dos meteorologistas. Entretanto, esses padrões eram conhecidos de um ponto de vista mais qualitativo. Assim, os resultados encontrados validam a metodologia proposta, a qual possibilita a mineração de dados em outros casos de teste mais específicos, de forma a tentar encontrar outros padrões desconhecidos que possam ser úteis para a Meteorologia e áreas afins.

Em paralelo, pretende-se validar a metodologia de agrupamento espaço-temporal de ocorrências de descargas em CAEs, empregada neste trabalho, visando sua operacionalização, no intuito de disponibilizar uma ferramenta de rastreamento de atividade convectiva para auxílio à previsão meteorológica, denominada *Cb-Trace*.

5. AGRADECIMENTOS

Os autores agradecem ao Met. Cesar A. A. Beneti (SIMEPAR), à Fundação Tecnológica SIMEPAR e ao RINDAT os dados de descargas elétricas atmosféricas utilizados neste trabalho, ao CPTEC/INPE o acesso aos dados de radiossondagem e de reanálise do NCEP, à FAPESP (projeto IPE, processo no. 1988/0105-5) os dados de radiossondagem deste projeto e ao CNPq o apoio financeiro fornecido aos projetos Cb-IPE e Electr (processos nos. 478707/2003-7 e 477819/03-6) e à bolsa de mestrado (processo no. 131384/2003-1).

6. REFERÊNCIAS BIBLIOGRÁFICAS

BAILEY, T.; GATRELL, A. **Interactive spatial data analysis.** Longman, 1995. 432 p.

BENETI, C. A. A.; ALVIM, E. L.; ANDRADE, S. M. G.; ASSUNÇÃO, L. A. R.; CAZETTA, A. F.; REIS, R. J. RIDAT – Rede Integrada de Detecção de Descargas Atmosféricas no Brasil: situação atual, aplicações e perspectivas. In: CONGRESSO BRASILEIRO DE METEOROLOGIA, 11., 2000, Rio de Janeiro. **Anais.** Rio de Janeiro: SBMET, 2000 (em CD-ROM).

CHEN, Z. **Data mining and uncertain reasoning: an integrated approach**. John Wiley & Sons, Inc., 2001.

COTTON, W. R.; ANTHES, R. A. **Storm and cloud dynamics**. San Diego Academic Press, 1989.

DOMINGUES, M. O., MENDES JR, O., CHAN, C. S., SA, L. D. A., MANZI, A. O. **Estado do céu durante o experimento IPE-2 do Projeto de Estudo da Camada Limite Superficial do Pantanal Sul Matogrossense**. São José dos Campos: Relatório Técnico INPE, 2002 (INPE-8861-NTC/438), em CD-ROM.

DOMINGUES, M. O.; MENDES JR, O.; CHAN, C. S.; BENETI, C. A. Atmospheric parameters related to lightning activity: events from dry season interdisciplinary Pantanal experiment in Brazil. In: 12th INTERNATIONAL CONFERENCE IN ATMOSPHERIC ELETRICITY, 2003, Versailles. **Proceedings of ICAE-2003.** Versailles: 2003.

DOMINGUES, M. O.; MENDES JR, O.; CHAN, C. S.; ALVALÁ, R. C. S.; ABREU DE SÁ, L. D.; MANZI, A. O. **Estado do céu durante o experimento IPE-3 do Projeto de Estudo da Camada Limite Superficial do Pantanal Sul Matogrossense**. São José dos Campos: Relatório Técnico INPE, 2003 (INPE-10045-NTC/359), em CD-ROM.

DOMINGUES, M. O.; MENDES JR, O.; CHAN, C. S.; SÁ, L. D. A.; MANZI, A. O. Análise das condições atmosféricas durante a 2a. Campanha do Experimento Interdisciplinar do Pantanal Sul Mato-Grossense. **Revista Brasileira de Meteorologia**, São Paulo, v.19, n.1, p.73-88, 2004.

DOSWELL, C. A.; RASMUSSEN, E. R. The effect of neglecting the virtual temperature correction on CAPE calculation. **Weather and Forecasting**, v.9, p.625-629, 1994.

DOSWELLL, C. A.; CARACENA, F.; MAGNANO, M. 1985: Temporal evolution of 700-500 mb lapse rate as a forecasting tool - a case study. 14TH CONFERENCE ON SEVERE LOCAL STORMS, 1985, Indianapolis. **Preprints.** Indianapolis: American Meteorological Society, 1985. p. 398-401.

EPANECHNIKOV, V. A. Nonparametric estimation of multidimensional probability density. **Theory of Probability and Its Applications**, v.14, n.2, p.153-158, 1969.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The KDD process for extracting useful knowledge from volumes of data. **Communications of the ACM**, v.39, n.11, pp.27-34, Nov. 1996.

FLAHAUL, B.; MOUCHART, M.; MARTIN, E. S.; THOMAS, I. The local spatial autocorrelation and the kernel method for identifying black zones - A comparative approach. **Accident Analysis and Prevention**, v. 35, n.6, p.991-1004, 2003.

GALWAY, J. G. The lifted index as a predictor of latent instability. **Bulletin of the American Meteorological Society**, v.29, n.37, p.528-529, 1956.

GEORGE, J. J. **Weather forecasting for aeronautics**. Academic Press, 1960. 673 p.

GRILLENZONI, C. Non-parametric smoothing of spatio-temporal point processes. **Journal of Statistical Planning and Inference**, v.33, n.2, p.25-36, 2004.

KALNAY, E.; KANAMITSU, M.; KISTLER, R.; COLLINS, W.; DEAVEN, D.; GANDIN, L.; IREDELL, M.; SAHA, S.; WHITE, G.; WOOLLEN, J; ZHU, J.; CHELLIAH, M.; EBISUZAKI, W.; HIGGINS, W.; JANOWIAK, J.; MO, K.; ROPELEWSKI, C.; WANG, J.; LEETMAA, A.; REYNOLDS, R.; JENNE, R.; JOSEPH, D. The NCEP/NCAR 40-year reanalysis project. **Bulletin of the American Meteorological Society**, v.77, n.3, p.437-471, 1996.

LEE, S. **Lecture notes for MECT2 nonparametric methods**, 2003.

MACGORMAN, D. R.; RUST, W. D. **The electrical nature of storms**. Oxford: Oxford University Press, 1998. 422 p.

MENDES JR., O.; DOMINGUES, M. O. Introdução à eletrodinâmica atmosférica. **Revista Brasileira de Ensino de Física**, v.24, n.1, p.3-19, mar. 2002.

MILLER, R.C. **Notes on analysis and severe storm forecasting procedures of the Air Force Global Weather Central. Tech. Rept. 200(R)**, Headquarters, Air Weather Service, USAF, 1972. 190 p.

MONCRIEFF, M. W.; GREEN, J. S. A. The propagation of steady convective overturning in shear. **Quarterly Journal of the Royal Meteorological Society**, v.98, n.3, p.336-352, 1972.

PAWLAK, Z. Rough sets. **International Journal of Computer and Information Sciences**, v.11, n.5, pp. 341-356, 1982.

PIATETSKY-SHAPIRO, G. Knowledge discovery in real databases, a report on the IJCAI-89 Workshop. **AI Magazine**, v.11, n.5, p.68-70, Jan. 1991, Special Issue.

POLITI, J.; DOMINGUES, M. O.; MENDES JR, O.; STEPHANY, S. Tracing atmospheric convective activity by means of data mining techniques. In: VII LATIN-AMERICAN CONFERENCE ON SPACE GEOPHYSICS, 2004, Atibaia. **Proceedings of the VII COLAGE**. Atibaia: 2004, p. 112.

POLITI, J.; STEPHANY, S.; DOMINGUES, M.O.; MENDES JR, O. A data mining methodology for tracing convective kernels from cloud-to-ground discharge and other atmospheric datasets. In: III CONFERÊNCIA CIENTÍFICA DO LBA - EXPERIMENTO DE GRANDE ESCALA DA BIOSFERA-ATMOSFERA NA AMAZÔNIA, 2004, Brasília. **Proceedings of the III LBA Conference**. Brasília: 2004, p. 27.

POLITI, J.; STEPHANY, S.; MENDES JR, O.; DOMINGUES, M. O. Implementação de um ambiente para mineração de dados aplicado ao estudo de núcleos convectivos. In: IV WORKSHOP DOS CURSOS DE COMPUTAÇÃO APLICADA DO INPE, 2004, S. José dos Campos. **Anais do IV WORCAP**. S. José dos Campos, 2004 (em CD-ROM).

POLITI, J. **Implementação de um ambiente para mineração de dados aplicada ao estudo de núcleos convectivos**. 2005. 146f. Dissertação (Mestrado em Computação Aplicada) – INPE, São José dos Campos, 2005 (INPE-14165-TDI/1082).

ØHRN, A. **Discernibility and rough sets in medicine: tools and applications**. Department of Computer and Information Science, Norwegian University of Science and Technology, 1999.

SILVERMAN, B. W. **Density estimation for statistics and data analysis (Monographs on Statistics and Applied Probability 26)**. London: Chapman and Hall, 1990.

UMAN, M. A. **The lightning discharge**. Florida: Academic Press, 1987, 377 p.

VOLLAND, H. **Atmospheric electrodynamics**. Berlin: Springer Verlag, 1984, 205 p.

WILLIAMS, E.; RENNO, N. An analysis of the conditional instability of the tropical atmosphere. **Monthly Weather Review**, v.121, n.1, p.21-36, 1993.