



Ministério da
**Ciência, Tecnologia
e Inovação**



sid.inpe.br/mtc-m19/2013/11.28.14.11-TDI

MONITORAMENTO E PREVISÃO DE ATIVIDADE CONVECTIVA USANDO ABORDAGENS DE MINERAÇÃO DE DADOS

Cesar Strauss

Tese de Doutorado do Curso de Pós-Graduação em Computação Aplicada, orientada pelos Drs. Stephan Stephany, e Marcelo Barbio Rosa aprovada em 29 de novembro de 2013.

URL do documento original:

<<http://urlib.net/8JMKD3MGP7W/3FACD25>>

INPE
São José dos Campos
2013

PUBLICADO POR:

Instituto Nacional de Pesquisas Espaciais - INPE

Gabinete do Diretor (GB)

Serviço de Informação e Documentação (SID)

Caixa Postal 515 - CEP 12.245-970

São José dos Campos - SP - Brasil

Tel.:(012) 3208-6923/6921

Fax: (012) 3208-6919

E-mail: pubtc@sid.inpe.br

CONSELHO DE EDITORAÇÃO E PRESERVAÇÃO DA PRODUÇÃO INTELLECTUAL DO INPE (RE/DIR-204):

Presidente:

Marciana Leite Ribeiro - Serviço de Informação e Documentação (SID)

Membros:

Dr. Antonio Fernando Bertachini de Almeida Prado - Coordenação Engenharia e Tecnologia Espacial (ETE)

Dr^a Inez Staciarini Batista - Coordenação Ciências Espaciais e Atmosféricas (CEA)

Dr. Gerald Jean Francis Banon - Coordenação Observação da Terra (OBT)

Dr. Germano de Souza Kienbaum - Centro de Tecnologias Especiais (CTE)

Dr. Manoel Alonso Gan - Centro de Previsão de Tempo e Estudos Climáticos (CPT)

Dr^a Maria do Carmo de Andrade Nono - Conselho de Pós-Graduação

Dr. Plínio Carlos Alvalá - Centro de Ciência do Sistema Terrestre (CST)

BIBLIOTECA DIGITAL:

Dr. Gerald Jean Francis Banon - Coordenação de Observação da Terra (OBT)

REVISÃO E NORMALIZAÇÃO DOCUMENTÁRIA:

Marciana Leite Ribeiro - Serviço de Informação e Documentação (SID)

Yolanda Ribeiro da Silva Souza - Serviço de Informação e Documentação (SID)

EDITORAÇÃO ELETRÔNICA:

Maria Tereza Smith de Brito - Serviço de Informação e Documentação (SID)

André Luis Dias Fernandes - Serviço de Informação e Documentação (SID)



Ministério da
**Ciência, Tecnologia
e Inovação**



sid.inpe.br/mtc-m19/2013/11.28.14.11-TDI

MONITORAMENTO E PREVISÃO DE ATIVIDADE CONVECTIVA USANDO ABORDAGENS DE MINERAÇÃO DE DADOS

Cesar Strauss

Tese de Doutorado do Curso de Pós-Graduação em Computação Aplicada, orientada pelos Drs. Stephan Stephany, e Marcelo Barbio Rosa aprovada em 29 de novembro de 2013.

URL do documento original:

<<http://urlib.net/8JMKD3MGP7W/3FACD25>>

INPE
São José dos Campos
2013

Dados Internacionais de Catalogação na Publicação (CIP)

Strauss, Cesar.

St82m Monitoramento e previsão de atividade convectiva usando abordagens de mineração de dados / Cesar Strauss. – São José dos Campos : INPE, 2013.

xx + 112 p. ; (sid.inpe.br/mtc-m19/2013/11.28.14.11-TDI)

Tese (Doutorado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2013.

Orientadores : Drs. Stephan Stephany, e Marcelo Barbio Rosa.

1. mineração de dados. 2. descargas elétricas atmosféricas. 3. sistemas convectivos. 4. previsão meteorológica. 5. estimação de densidade I.Título.

CDU 681.3.01

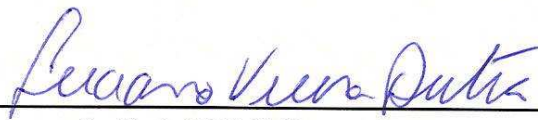


Esta obra foi licenciada sob uma Licença [Creative Commons Atribuição-NãoComercial 3.0 Não Adaptada](https://creativecommons.org/licenses/by-nc/3.0/).

This work is licensed under a [Creative Commons Attribution-NonCommercial 3.0 Unported License](https://creativecommons.org/licenses/by-nc/3.0/).

Aprovado (a) pela Banca Examinadora
em cumprimento ao requisito exigido para
obtenção do Título de **Doutor(a)** em
Computação Aplicada

Dr. Luciano Vieira Dutra



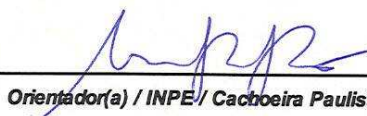
Presidente / INPE / SJC Campos - SP

Dr. Stephan Stephany



Orientador(a) / INPE / SJC Campos - SP

Dr. Marcelo Barbio Rosa




Orientador(a) / INPE / Cachoeira Paulista - SP

Dr. Rodrigo Arnaldo Scarpel



Convidado(a) / ITA / São José dos Campos - SP

Dr. Tércio Ambrizzi



Convidado(a) / IAG/USP / São Paulo - SP

Este trabalho foi aprovado por:

maioria simples

unanimidade

Aluno (a): **Cesar Strauss**

São José dos Campos, 29 de Novembro de 2013

RESUMO

Neste trabalho são propostas algumas abordagens de mineração de dados para o monitoramento e previsão de atividade convectiva atmosférica. Os dados meteorológicos são referentes a descargas elétricas atmosféricas, radares meteorológicos e modelos numéricos de previsão de tempo. Uma primeira abordagem é a geração de campos de densidade de ocorrência de descargas para identificar as regiões mais densas, correspondentes a células de tempestade eletricamente ativas, implementada pelo software EDDA. Foi desenvolvida uma funcionalidade extra deste software, a qual também executa o agrupamento espaço-temporal das descargas, identificando assim células elétricas com mais precisão. Testes em ambiente operacional demonstraram que esses campos de densidade de descargas podem monitorar de maneira mais efetiva a atividade convectiva em comparação a outras técnicas correntes que empregam imagens de satélites meteorológicos. Naturalmente, os radares meteorológicos são mais efetivos, porém sua cobertura espacial é muito limitada no território brasileiro. A segunda abordagem baseia-se na identificação de padrões associados à atividade convectiva em saídas de um modelo numérico de previsão de tempo, de forma a prever a ocorrência e distribuição de chuva forte e convectiva a partir das previsões desse modelo. Essa abordagem foi implementada no software EPPA, ainda em desenvolvimento, que utiliza uma técnica de aprendizado de máquina, uma árvore de decisão. Testes preliminares demonstraram que esse software tem potencial para melhorar a previsão de chuva, uma vez que os modelos numéricos são sabidamente imprecisos nesse aspecto.

MONITORING AND PREDICTION OF CONVECTIVE EVENTS USING DATA MINING APPROACHES

ABSTRACT

In this work, some data mining approaches are proposed for the monitoring and prediction of atmospheric convective activity. The related meteorological data include atmospheric electrical discharges, weather radar images and numerical weather forecast outputs. A first approach is the generation of fields of density of occurrence of discharges in order to identify the denser areas that correspond to electrically active storm cells. This approach is implemented by the EDDA software. An extra functionality of this software performs the spatio-temporal clustering of the electrical discharges allowing a more precise identification of the electrical cells. Tests performed in an operation environment show that these fields of density of discharges are more effective for the monitoring of convective activity in comparison to other current techniques that use satellite images. Naturally, weather radar are more effective, but their spatial coverage is very limited in the Brazilian territory. The second approach is based on the identification of patterns associated to convective activity in the outputs of a numerical weather forecast model, in order to predict the occurrence and distribution of intense and convective rain from the model forecasts. This approach was implemented in the EPPA software that is being developed. It employs a machine learning technique, a decision tree. Preliminary tests show that this software may potentially improve rain prediction, since numerical models are known to be imprecise for such task.

LISTA DE FIGURAS

	<u>Pág.</u>
2.1 Estrutura elétrica de uma nuvem de tempestade	10
2.2 Localização dos sensores da rede RINDAT	11
2.3 Exemplo de visualização de campo de densidade de ocorrência de descargas elétricas atmosféricas, gerado pelo software EDDA.	13
2.4 Exemplo de imagem CAPPI compondo dados dos radares meteorológicos de Bauru e Presidente Prudente (os círculos têm 240 km de raio)	15
3.1 Subdivisão de uma grade regular pelo algoritmo diamante-quadrado. O valor do campo nos círculos pretos é calculado em função dos valores dos círculos brancos adjacentes.	27
3.2 O campo de densidade aleatório gerado método fractal para teste (acima), curvas de nível correspondente (meio) e conjunto aleatório de eventos gerado a partir desse campo (em baixo).	28
3.3 Campos de densidade estimados usando os vários esquemas de ajuste de largura de janela a partir do conjunto aleatória de eventos considerado.	29
3.4 Fluxo do processo de descoberta de conhecimento em banco de dados (KDD).	31
3.5 Ilustração do método de subida de encosta, aplicado sobre o campo de densidade de instâncias, que é utilizado no método de agrupamento DEN-CLUE 2.0. Fonte: Hinneburg e Gabriel (2007)	33
3.6 Exemplo de árvore de decisão para verificar ocorrência (TRUE) ou ausência (FALSE) de precipitação forte ou convectiva e forte a partir de índices de instabilidade e variáveis atmosféricas	34
3.7 Série temporal dos valores acumulados de descargas NS em 30 min e a massa precipitada (chuva) para ano de 2009 dentro da área A de 2.500 km ² próxima Bauru (curvas suavizadas por um filtro gaussiano unidimensional).	39
3.8 Série temporal dos valores acumulados de descargas NS em 30 min e a massa precipitada (chuva) para o mês de setembro de 2009 dentro da área A de 2.500 km ² próxima a Bauru (curvas suavizadas por um filtro gaussiano unidimensional).	40

3.9	Série temporal dos valores acumulados de descargas NS em 30 min e a massa precipitada (chuva) para alguns dias no mês de setembro de 2009 dentro da área A de 2.500 km ² próxima a Bauru durante uma tempestade em particular (curvas suavizadas por um filtro gaussiano unidimensional).	41
4.1	Fluxograma das implementações operacionais do software de EDDA (à esquerda) e o mesmo software com agrupamento de descargas (à direita).	46
4.2	Um exemplo da evolução do agrupamentos de descargas de um intervalo de tempo para outro, de acordo com o critério de fusão/separação adotado.	48
4.3	Comparação entre a grade de radar (0,02°) e a grade do modelo ETA (0,05°).	51
5.1	Imagem de densidade de descargas (EDDA) sobreposta à imagem GOES realçada.	56
5.2	Imagem de densidade de descargas (EDDA) sobreposta aos sistemas convectivos (ForTraCC) e às estações do INMET.	57
5.3	Imagem de densidade de descargas (EDDA) sobreposta à imagem do radar de São Roque (CAPPI 3 km) e às estações do INMET.	57
5.4	Imagem de densidade de descargas (EDDA) sobreposta à precipitação instantânea estimada por satélite (Hidroestimador), aos sistemas convectivos e às estações INMET.	58
5.5	(Superior) Imagens do satélite GOES-12 no canal infravermelho (canal 4 = 3,8–4,0 μm), em 16/01/2010 às 20:00 UTC para o evento A (à esquerda) e em 19/01/2010 às 22:00 UTC para evento B (à direita). (Inferior) Previsão do modelo atmosférico de circulação geral (AGCM), do CPTEC para o evento A em 16/01/2010 18:00 UTC (à esquerda) e para o evento B em 19/01/2010 18:00 UTC (à direita) mostrando pressão na superfície (sombreada) e o padrão de circulação em 250 hPa (linhas de contorno).	61
5.6	Topo: Refletividade do radar (dBZ) do evento A em 16/01/2010 às 20:10 UTC. Meio e inferior: células eletricamente ativas do evento A detectado unicamente pela estimação de densidade (meio) e com o agrupamento (parte inferior). Tons de cinza correspondem a precipitação estratiforme, e tons escuros, a precipitação convectiva. Cada célula é mostrada por meio de seus contornos em um limiar de densidade determinado.	63

5.7	Topo: Refletividade do radar (dBZ) do evento B em 19/01/2010 às 23:00 UTC. Meio e inferior: células eletricamente ativas do evento B detectado unicamente pela estimação de densidade (meio) e com o agrupamento (inferior). Tons de cinza correspondem a precipitação estratiforme e tons escuros, a precipitação convectiva. Cada célula é mostrada por meio de seus contornos em um limiar de densidade determinado.	64
5.8	Rastros no solo dos centroides das células eletricamente ativas (cinza) e células de precipitação (preto) no evento A (16/01/2010 de 18:30 para 20:30 UTC), destacando as células #1 para #3.	66
5.9	Rastros no solo de células eletricamente ativas (cinza) e células de precipitação (preto) no evento B (19/01/2010 de 22:00 a 23:30 UTC), destacando as células #1 para #5.	67
5.10	Evolução temporal do número de descargas NS acumulado em 10 min e a precipitação média para a célula #1 do evento A (topo) e células #3 do evento B (inferior) que passou por cima de Bauru.	68
5.11	Imagens meteorológicas relativas às três tempestades consideradas: evento de 14/01/2010 às 18:00 UTC (esquerda), evento de 20/01/2010 às 00:00 UTC (centro) e evento de 26/01/2010 às 18:00 UTC (direita). Na fileira superior aparecem as imagens do satélite GOES-12 no banda infravermelha, na fileira do meio os correspondentes campos de pressão e linhas de corrente (LC) e, na fileira inferior, as refletividades (dBZ) medidos pelos radares meteorológicos.	73
5.12	Comparação das classificações efetuadas pela ferramenta de previsão objetiva do CPTEC (caso 1), à esquerda, e pelo software EPPA para previsão de ocorrência de precipitação, à direita, para a tempestade do dia 14/01/2010 às 18:00 UTC, considerando-se quatro diferentes previsões.	75
5.13	Comparação das classificações efetuadas pela ferramenta de previsão objetiva do CPTEC (caso 1), à esquerda, e pelo software EPPA para previsão de ocorrência de precipitação, à direita, para a tempestade do dia 20/01/2010 às 00:00 UTC, considerando-se quatro diferentes previsões.	76
5.14	Comparação das classificações efetuadas pela ferramenta de previsão objetiva do CPTEC (caso 1), à esquerda, e pelo software EPPA para previsão de ocorrência de precipitação, à direita, para a tempestade do dia 26/01/2010 às 18:00 UTC, considerando-se quatro diferentes previsões.	77
5.15	Gráfico do número de falsos negativos em função dos falsos positivos para o conjunto de dados de validação do modelo ETA 5 km correspondente a janeiro de 2010, considerando-se uma vizinhança de 3×3 pixels. As ordenadas variam de 0 a 10000 e as abcissas, de 0 a 12000.	82

5.16	Previsão da árvore de decisão (EPPA) relativa à ocorrência de precipitação forte ou convectiva para a tempestade do dia 14/01/2010. A sequência de imagens corresponde aos horários das 15, 18 e 21 UTC, e 0 UTC do dia seguinte.	83
5.17	Previsão da árvore de decisão (EPPA) relativa à ocorrência de precipitação forte ou convectiva para a tempestade do dia 19/01/2010. A sequência de imagens corresponde aos horários das 15, 18 e 21 UTC, e 0 UTC do dia seguinte.	84
5.18	Previsão da árvore de decisão (EPPA) relativa à ocorrência de precipitação forte ou convectiva para a tempestade do dia 26/01/2010. A sequência de imagens corresponde aos horários das 16, 18 e 21 e 23 UTC.	85

LISTA DE TABELAS

	<u>Pág.</u>
3.1 Erro médio quadrático integrado (MISE) dos vários esquemas de janela fixa e adaptativa para o conjunto de eventos considerado.	27
5.1 Média ponderada das correlações entre as 3 células eletricamente ativas do evento A às 20:10 UTC e a chuva convectiva correspondente para diferentes valores de deslocamento (zero, até 1 pixel, ou até 3 pixels). Os três valores referem-se à correlação entre as células e a chuva observada na iteração anterior (20:00 UTC), a corrente (20:10 UTC) e o subsequente (20:20 UTC), respectivamente, para cada célula. As melhores correlações são destacadas.	65
5.2 Média ponderada das correlações entre as 5 células eletricamente ativas do evento B às 23:00 UTC e a correspondente chuva convectiva para diferentes valores de deslocamento espacial (nenhuma mudança, até 1 ou 3 pixels). Os três valores referem-se à correlação entre as células e a chuva observada na iteração anterior (22:50 UTC), a corrente (23:00 UTC) e o subsequente (23:10 UTC), respectivamente, para cada célula. As melhores correlações são destacadas.	67
5.3 Desempenho médio da Ferramenta Objetiva de Previsão do Tempo do CPTEC/INPE para previsão de pancadas de chuva com trovoadas (tipo 1). O valores referem-se a cada conjunto de quatro previsões do ETA 20 km considerando diversas vizinhanças.	71
5.4 Desempenho médio da Ferramenta Objetiva de Previsão do Tempo do CPTEC/INPE para previsão de tempestades (tipo 2). O valores referem-se a cada conjunto de quatro previsões do ETA 20 km considerando diversas vizinhanças.	71
5.5 Desempenho médio da árvore de decisão (EPPA) para previsão de ocorrência de precipitação, treinando e testando com dados de análise e previsão do modelo ETA 20 km.	71
5.6 Desempenho médio da árvore de decisão (EPPA) para previsão de ocorrência de precipitação, treinando com a análise (00 UTC) e previsões de 06 – 18 UTC, e testando com as demais previsões do modelo ETA 20 km.	71

5.7	Comparação dos desempenhos da Ferramenta Objetiva de Previsão do Tempo do CPTEC/INPE e do software EPPA na previsão de ocorrência de precipitação para a tempestade do dia 14/01/2010 às 18:00 UTC para quatro previsões diferentes.	74
5.8	Comparação dos desempenhos da Ferramenta Objetiva de previsão do Tempo do CPTEC/INPE e do software EPPA na previsão de ocorrência de precipitação para a tempestade do dia 20/01/2010 às 00:00 UTC para quatro previsões diferentes.	74
5.9	Comparação dos desempenhos da Ferramenta Objetiva de Previsão do Tempo do CPTEC/INPE e do software EPPA na previsão de ocorrência de precipitação para a tempestade do dia 26/01/2010 às 18:00 UTC para quatro previsões diferentes.	78
5.10	Desempenho médio da árvore de decisão (EPPA) para previsão de ocorrência de chuva forte ou convectiva, treinando com a análise (00 UTC) e previsões de 06 – 18 UTC, e testando com as demais previsões do modelo ETA 20 km.	78
5.11	Desempenho médio da árvore de decisão (EPPA) para previsão de ocorrência de precipitação forte e convectiva, treinando e testando com dados de análise e previsão do modelo ETA 5 km, considerando-se ou não vizinhança temporal de ± 30 min e vizinhança espacial de 3×3 pixels.	81
5.12	Desempenho de classificação da árvore de decisão (EPPA) para previsão de ocorrência de precipitação forte e convectiva, sem vizinhança espacial ou temporal, para a tempestade do dia 14/01/2010, para as quatro previsões consideradas.	83
5.13	Desempenho de classificação da árvore de decisão (EPPA) para previsão de ocorrência de precipitação forte e convectiva, sem vizinhança espacial ou temporal, para a tempestade do dia 19/01/2010.	84
5.14	Desempenho de classificação da árvore de decisão (EPPA) para previsão de ocorrência de precipitação forte e convectiva, sem vizinhança espacial ou temporal, para a tempestade do dia 26/01/2010.	85
5.15	Desempenho médio da árvore de decisão (EPPA) para previsão de ocorrência de chuva forte ou convectiva, com dados do modelo ETA 5 km para o verão de 2009/2010, utilizando diversos tipos de treinamento da árvore de decisão.	88
5.16	Desempenho médio da árvore de decisão (EPPA) para previsão de ocorrência de chuva forte ou convectiva, com dados do modelo ETA 5 km para o verão de 2010/2011, utilizando diversos tipos de treinamento da árvore de decisão.	88

5.17	Desempenho médio da árvore de decisão (EPPA) para previsão de ocorrência de chuva forte ou convectiva, com dados do modelo ETA 5 km para o verão de 2010/2011, utilizando <i>hold out</i> cronológico do tipo II, para diferentes limiares de impureza.	89
5.18	Importância relativa (em %) de cada atributo conforme observado no treinamento da melhor árvore (<i>hold out</i> cronológico do tipo II, com limiar de impureza de 5%) com os dados do modelo ETA 5 km para o verão de 2010/2011.	89
5.19	Matriz de confusão para o conjunto de teste (em %) da melhor árvore obtida (<i>hold out</i> cronológico do tipo II, com limiar de impureza de 5%) com os dados do modelo ETA 5 km para o verão de 2010/2011. As células referentes aos falsos positivos e negativos (diagonal secundária) foram quebradas de maneira a apresentar a proporção de falsos positivos e de falsos negativos que foram reclassificados como verdadeiros positivos devido à vizinhança relaxada.	89
5.20	Simplificação da tabela anterior, correspondente à melhor árvore obtida (<i>hold out</i> cronológico do tipo II, com limiar de impureza de 5%) com os dados do modelo ETA 5 km para o verão de 2010/2011.	89

LISTA DE ABREVIATURAS E SIGLAS

BrasilDAT	–	Sistema Brasileiro de Detecção de Descargas Atmosféricas
CAPPI	–	Constant Altitude Plan Position Indicator
CEMADEN	–	Centro Nacional de Monitoramento e Alertas de Desastres Naturais órgão do Ministério da Ciência, Tecnologia e Inovação
CNPq	–	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CPTEC	–	Centro de Previsão do Tempo e Estudos Climáticos (INPE)
DECEA	–	Departamento de Controle do Espaço Aéreo do Ministério da Aeronáutica
DSA	–	Divisão de Satélites e Sistemas Ambientais (CPTEC/INPE)
ECWMF	–	European Centre for Medium-Range Weather Forecasts
EDDA	–	Estimador de Densidade de Descargas Atmosféricas
EPPA	–	Estimador de Precipitação usando dados de Previsão e Análise de modelo numérico
FAR	–	False Alarm Ratio
FINEP	–	Agência Brasileira da Inovação
GOES	–	Geostationary Operational Environmental Satellite
IA	–	Inteligência Artificial
IN	–	Intra-nuvem
INPE	–	Instituto Nacional de Pesquisas Espaciais
INMET	–	Instituto Nacional de Meteorologia
IPMet	–	Instituto de Pesquisas Meteorológicas (UNESP)
JAXA	–	Agência espacial japonesa.
LAC	–	Laboratório Associado de Computação e Matemática Aplicada (INPE)
MOS	–	Model Output Statistics
NOAA	–	National Oceanic and Atmospheric Administration.
NASA	–	National Aerospace Administration, agência espacial americana
NDVI	–	Normalized Difference Vegetation Index
NEN	–	Nível de Empuxo Neutro
NCC	–	Nível de Convecção Convectiva
NN	–	Nuvem-nuvem
NS	–	Nuvem-solo
POD	–	Probability of Detection
RINDAT	–	Rede Integrada Nacional de Detecção de Descargas Atmosféricas
RLR	–	Rainfall–Lightning Ratio
SCM	–	Sistemas Convectivos de Mesoescala.
SUP	–	Nível de Superfície
TRMM	–	Tropical Rainfall Measuring Mission
UALF	–	Universal ASCII Lightning Format
WDSS-II	–	Warning Decision Support System – Integrated Information
WV-IR	–	Índice Water Vapor - InfraRed

SUMÁRIO

	<u>Pág.</u>
1 INTRODUÇÃO	1
1.1 Revisão Bibliográfica	2
1.2 Contribuições propostas neste trabalho	6
2 DADOS METEOROLÓGICOS	9
2.1 Dados de descargas elétricas atmosféricas	9
2.2 Dados de radar meteorológico	12
2.3 Dados de modelos numéricos	16
3 ALGUNS MÉTODOS RELACIONADOS A ESTE TRABALHO	23
3.1 Estimação de densidade	23
3.1.1 Avaliação do uso de largura de janela fixa ou adaptativa	26
3.2 Métodos de agrupamento e classificação	29
3.2.1 O método de agrupamento DENCLUE	31
3.2.2 Árvores de decisão	33
3.3 Ferramenta Objetiva de Auxílio à Previsão do Tempo	36
3.4 Análise da correlação entre número descargas NS e chuva convectiva . . .	38
3.5 Estimação da precipitação acumulada a partir de descargas elétricas at- mosféricas	40
3.6 Critério de identificação de chuva convectiva em imagens de radar	42
4 OS SOFTWARES EDDA E EPPA	43
4.1 O software EDDA	43
4.2 O software EDDA com agrupamento espaço-temporal de descargas	45
4.3 O software EDDA com estimação de precipitação	49
4.4 O software EPPA	49
5 RESULTADOS	53
5.1 Uso do software EDDA para monitoramento de atividade convectiva . . .	54
5.2 Uso da ferramenta EDDA com agrupamento de descargas para monito- ramento e rastreamento de atividade convectiva	58
5.2.1 Análise sinótica das tempestades selecionadas	59

5.2.2	Análise da correlação entre descargas e atividade convectiva para os eventos selecionados	60
5.2.3	Análise de correlação entre células eletricamente ativas e células de chuva para os eventos selecionados	62
5.2.4	Comparação das trajetórias das células eletricamente ativas e das células de chuva convectiva para os eventos selecionados	68
5.2.5	Considerações relativas à evolução temporal para os eventos selecionados	69
5.3	O software EPPA para previsão de ocorrência de precipitação	69
5.3.1	Previsão de ocorrência de precipitação com dados do modelo ETA 20 km	70
5.3.2	Previsão de ocorrência de precipitação com dados do modelo ETA 5 km	79
6	CONCLUSÕES E COMENTÁRIOS FINAIS	93
	REFERÊNCIAS BIBLIOGRÁFICAS	97
	ANEXO A - ARTIGOS PUBLICADOS RELACIONADOS À TESE	107

1 INTRODUÇÃO

O monitoramento e prevenção de desastres naturais vêm impulsionando a pesquisa de novos softwares para auxiliar os órgãos governamentais a emitir de maneira rápida e precisa os alertas para a defesa civil. Dentre esses órgãos, este trabalho está relacionado com o Centro de Previsão de Tempo e Estudos Climáticos do Instituto Nacional de Pesquisas Espaciais (CPTEC/INPE) e o Centro Nacional de Monitoramento e Alertas de Desastres Naturais (CEMADEN).

Dois tipos de abordagens podem ser considerados na tomada de decisão em Meteorologia: o monitoramento e a previsão meteorológica. O monitoramento sumariza a saída de sensores meteorológicos em imagens de múltiplas camadas que, sendo atualizadas em tempo real, permitem aos especialistas estarem sempre a par do estado atual da situação. Permite também extrapolar a tendência atual para as próximas horas, no escopo da previsão de curtíssimo prazo (*nowcasting*). A previsão meteorológica utiliza simulação numérica para prever eventos meteorológicos futuros a partir de simulação numérica utilizando modelos atmosféricos.

O primeiro objetivo deste trabalho é desenvolver e implementar uma metodologia para monitoramento de atividade convectiva em tempo quase real a partir de dados de descargas elétricas atmosféricas. O segundo objetivo é a previsão de atividade convectiva a partir de dados de modelo numérico de previsão do tempo, mais especificamente, a previsão da precipitação convectiva. Isso requer a identificação de padrões associados à atividade convectiva nas previsões do modelo, padrões estes que são compostos por variáveis atmosféricas e índices de instabilidade atmosféricos.

Tendo em vista esses objetivos, foram desenvolvidos alguns softwares, bem como alguns métodos relacionados, visando contribuir tanto no monitoramento como na previsão de atividade convectiva severa. O software EDDA, Estimador de Densidade de Descargas Atmosféricas, permite a visualização, identificação e monitoramento de células convectivas a partir da densidade de descargas elétricas atmosférica nuvem-solo (NS) geradas no interior de nuvens de tempestade. Esse software utiliza dados coletados por uma rede ampla de sensores de descargas elétricas, a Rede Integrada Nacional de Detecção de Descargas Atmosféricas (RINDAT). Dados de descargas proporcionam uma melhor resolução espacial comparados a dados de pluviômetros e melhor cobertura espacial, comparados aos dados de radares meteorológicos no território brasileiro. O software EDDA permite ainda estimar espacialmente a precipitação acumulada, importante para os alertas de deslizamento e enchentes. Por outro lado, o software EPPA, Estimador de Precipitação usando dados de Previsão e

Análise de modelo numérico, objetiva a previsão de precipitação a partir do modelo numérico de previsão do tempo ETA. Ao longo deste texto, os termos precipitação e chuva serão usados com o mesmo significado.

A pesquisa relacionada a esta tese iniciou-se anos atrás com dois projetos envolvendo o Laboratório Associado de Computação e Matemática Aplicada (LAC/INPE) e também o CPTEC/INPE. O primeiro foi um projeto do Edital Universal do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), “CB-MINING: Mineração de dados associados a sistemas convectivos” (duração 2006–2010) e o segundo, um projeto financiado pela Agência Brasileira da Inovação (FINEP), “Tempestades: desenvolvimento de um sistema dinamicamente adaptativo para produção de alertas para a região Sul/Sudeste (ADAPT)” (duração 2007-2009), mais especificamente a Meta 2 – “Mineração de dados para identificação de condições favoráveis à gênese e evolução de tempestades” (um dos orientadores desta tese foi coordenador de ambos os projetos).

A presente tese tem a seguinte estrutura: o Capítulo 2 trata dos dados meteorológicos. O Capítulo 3 detalha alguns métodos já existentes que foram utilizados nos softwares propostos, enquanto que o Capítulo 4 aborda os softwares EDDA e EPPA. Finalmente, o Capítulo 5 apresenta os resultados obtidos e o Capítulo 6, as conclusões e comentários finais.

1.1 Revisão Bibliográfica

Uma hipótese básica que motivou o desenvolvimento do software EDDA é a existência de uma correlação entre descargas NS e atividade convectiva. Além deste tópico, esta revisão apresenta trabalhos relativos à aplicação de técnicas de mineração de dados em Meteorologia.

Ao longo dos anos, vários trabalhos procuraram uma correlação entre descargas elétricas atmosféricas e atividade convectiva. Por exemplo, [Carey e Rutledge \(2000\)](#) encontraram uma correlação forte entre a precipitação na fase de gelo e a produção de descargas na convecção observada em ilhas tropicais. [Machado et al. \(2009\)](#) encontraram uma forte correlação entre a probabilidade de ocorrência de descargas e o índice WV-IR (vapor d’água menos infravermelho) baseada em imagens de satélite *Geostationary Operational Environmental Satellite* (GOES) que mostram a temperatura do topo das nuvens. [Petersen et al. \(1996\)](#) examinaram o perfil vertical da refletividade (fração da energia emitida por um radar meteorológico refletida por água ou gelo na atmosfera) no Oceano Pacífico Ocidental, determinando que a maior

parte da atividade convectiva não produz nenhuma descarga NS. Entretanto, quando descargas NS estão presentes, a atividade convectiva realmente parece produzir uma precipitação mais forte, comparada à atividade convectiva sem descargas. Petersen e Rutledge (2001) encontraram uma grande variabilidade no regime convectivo ao longo de diversas regiões tropicais, ao examinar o perfil vertical de refletividade e a ocorrência de descargas medidas pelo satélite *Tropical Rainfall Measuring Mission* (TRMM). Mesmo considerando essa variabilidade, os autores encontraram uma correlação linear entre densidade de descargas e conteúdo de gelo precipitante.

Lang e Rutledge (2011) perceberam um padrão em que tempestades que produzem descargas elétricas NS positivas (a carga transferida da nuvem para o solo é positiva) apresentaram mais precipitação que tempestades com poucas descargas NS no total. Siingh et al. (2013) encontraram uma boa correlação entre raios (formados por um grupo de descargas sucessivas) e precipitação convectiva para duas regiões no sul-sudeste da Ásia com diversos usos da terra e climas. Cecil et al. (2005) examinaram um banco de dados de eventos de precipitação derivados de dados do satélite TRMM. Eles descobriram uma diferença significativa entre terra firme e oceano: enquanto 42% dos eventos em terra firme apresentaram raios quando a refletividade é 30 dBZ na altura em que a temperatura é -20°C , apenas 13% dos eventos no oceano apresentam raios nessas condições.

Barnolas et al. (2008) calcularam um coeficiente de correlação de Pearson de 0,61 entre a precipitação medida por pluviômetros e a taxa de raios NS num raio de 6 km desses pluviômetros num evento de inundação ocorrido na Catalunha, Espanha. Esta correlação aumenta para 0,78 quando se consideram apenas dados com altas taxas de descargas (acima de mil descargas em 30 min). Michaelides et al. (2010) analisaram 19 eventos chuvosos no Chipre e relacionaram a precipitação medida por uma rede de pluviômetros com a atividade de descargas correspondente. Eles encontraram que, para um raio de 10 km, 42% dos eventos tem a correlação de Pearson entre 0,80 e 1,00. Assumindo a correlação entre descargas NS e precipitação convectiva, Tapia et al. (1998) estimaram uma razão *Rainfall-Lightning Ratio* (RLR) para estimar a massa de água precipitada numa tempestade a partir de dados de descargas. Garcia et al. (2013) ajustaram uma função não-linear para obter a massa precipitada em função do número de descargas. Numa outra abordagem, Mosier et al. (2011) propuseram um critério para prever descargas a partir de dados de radar, e Harats et al. (2010) desenvolveram uma versão modificada do índice de instabilidade K para prever a probabilidade de descargas com base no modelo atmosférico do *European Centre for Medium-Range Weather Forecasts* (ECMWF).

Mattos e Machado (2011) analisaram o ciclo de vida de Sistemas Convectivos de Mesoescala (SCM) sobre o estado de São Paulo, usando dados de satélite nas bandas infravermelha e de microondas, assim como dados de descargas. Utilizando como métrica o coeficiente de correlação de Pearson, constataram que a ocorrência de descargas NS estava bem correlacionada ao tamanho do sistema convectivo (0,96), à altura de topo das nuvens (0,84), ao conteúdo integrado de gelo (0,86) e ao tamanho das partículas precipitantes (0,90). Na média, tempestades elétricas, ou seja, com ocorrência de descargas, têm uma duração maior e uma área maior que tempestades comuns. O máximo de densidade de descargas ocorre no início do ciclo do SCM, enquanto que a taxa de ocorrência de descargas alcança um máximo durante a fase de crescimento, perto da maturação do SCM. A correlação entre descargas e sistemas convectivos na cidade de São Paulo é enfocada em Oliveira e Mattos (2011), considerando um quadrado de 1° de lado, e concluindo que durante o verão há mais chuva e ocorrência de descargas, porém com descargas NS menos intensas, enquanto que no inverno há menos chuva e menos ocorrências, porém estas são mais intensas. O trabalho atribui a baixa correlação direta entre precipitação e descargas a uma defasagem temporal entre esses fenômenos. Beneti et al. (2012) demonstram haver correlação entre dados de radar e dados de descargas NS para o ciclo diurno de SCMs no Sudeste brasileiro.

Uma outra linha de pesquisa estuda o chamado “salto de descargas” (*lightning jump*), um aumento repentino na taxa de descargas que marca o início de tempo severo (granizo, ventos destruidores e tornados), como no estudo de Williams et al. (1999) e de Gatlin e Goodman (2010). Ambos trabalhos empregam dados de descargas NS e intra-nuvem (IN) e/ou dados de descargas 3D e radar meteorológico. Pineda et al. (2011) não encontraram um salto de descargas NS, mas somente de descargas IN. Schultz et al. (2011) encontraram uma melhor probabilidade de detecção de tempo severo (79% versus 66%) e uma menor taxa de falsos alarmes (36% versus 54%) ao usar descargas totais (NS mais IN) em vez de apenas taxa de descargas NS, usando um algoritmo para detectar saltos de descargas.

No caso do acompanhamento de células de tempestade eletricamente ativas, Lakshmanan et al. (2004) descreveram o ambiente *Warning Decision Support System - Integrated Information* (WDSS-II) que integra dados de radar meteorológico, satélite e descargas. O software EDDA aqui apresentado é semelhante ao software de visualização de descargas elétricas atmosféricas desse ambiente, que fornece monitoramento e previsão da evolução de células eletricamente ativas (Kohn et al., 2011). Outro trabalho que emprega a densidade de descargas é o de Betz et al. (2008), para

o rastreamento de células elétricas usando dados fornecidos pela rede europeia de detecção de descargas (LINET).

Este trabalho também aborda a previsão de chuvas fortes ou convectivas a partir de variáveis e índices obtidos de modelos numéricos de previsão do tempo. Esses modelos tem pouca eficácia em prever precipitação de forma confiável. Por exemplo, [Liguori et al. \(2012\)](#) constataram que o modelo MM5 tem melhor desempenho na previsão de chuva estratiforme do que da convectiva. Adicionalmente, [Ebert et al. \(2007\)](#) alegam que imagens infravermelhas de satélite são melhores que modelos numéricos para detectar chuva convectiva.

Assim como neste trabalho, técnicas de mineração de dados foram usadas em Meteorologia com diversas finalidades. Por exemplo, [Khan et al. \(2007\)](#) analisaram a variabilidade espacial e temporal de extremos de precipitação na América do Sul com base em dados de precipitação diários disponíveis numa grade de $2,5^\circ$ entre 1940 e 2004. [Gurgel et al. \(2003\)](#) analisaram a variabilidade do Índice de Vegetação por Diferença Normalizada (NDVI) sobre o Brasil, utilizando a análise de agrupamentos. Constataram, por exemplo, que a ocorrência de El Niño, independentemente de sua intensidade, afeta distintamente os vários tipos de vegetação. [Hoffman et al. \(2005\)](#) agruparam vetores de variáveis atmosféricas geradas por um modelo climático global entre 2000 e 2008. Cada grupo resultante foi associado a um diferente regime climático. Assim, confirmaram a tendência de aumento da desertificação global e o aquecimento da Antártica e da Groenlândia. No Brasil, [Vila et al. \(2008\)](#) desenvolveram o algoritmo “Previsão a Curto Prazo da Evolução de Sistemas Convectivos” ou *Forecast and Tracking the Evolution of Cloud Clusters* (ForTraCC) para rastrear sistemas convectivos de mesoescala usando imagens de satélites geo-estacionários GOES. Os agrupamentos de nuvens resultantes são definidos por meio de um limiar de temperatura de brilho de topo de nuvem e rastreados através da sobreposição de agrupamentos em intervalos sucessivos de tempo.

Mais recentemente, [Pessoa et al. \(2012\)](#) desenvolveram métodos de classificação supervisionada, com o objetivo de previsão de eventos convectivos severos, usando dados de um modelo numérico de previsão do tempo e dados de descargas elétricas atmosféricas. [Dolif e Nobre \(2012\)](#) usaram uma rede neural para previsão de ocorrência de precipitação intensa no Rio de Janeiro utilizando dados de pluviômetros. [Strauss et al. \(2012\)](#) utilizaram árvore de decisão para prever a ocorrência de chuva a partir de índices de instabilidade e variáveis atmosféricas provenientes do modelo numérico ETA. [Lima e Stephany \(2013a\)](#) propuseram um novo método de

agrupamento e classificação baseado na frequência de ocorrência de padrões, com o objetivo de encontrar padrões associados a tempo severo nas previsões do modelo numérico ETA, assumindo que altas densidades de descargas elétricas sejam indicativas de atividade convectiva. Uma abordagem similar foi proposta por Lima e Stephany (2013b), porém usando uma rede neural como classificador.

No tocante à previsão de chuva, uma abordagem bastante utilizada é o método *Model Output Statistics* (MOS) (GLAHN; LOWRY, 1972) que consiste em aplicar regressão linear em séries históricas de precipitação medidas em estações meteorológicas nos Estados Unidos, em função de variáveis numéricas do modelo meteorológico. A probabilidade de chuva futura é estimada para cada estação meteorológica individual. Num trabalho mais recente, Glahn et al. (2009) estimaram a precipitação numa grade a partir de regressão de variáveis estimadas por assimilação de dados, como por exemplo temperatura de superfície. Finalmente, Charba e Samplatsky (2011) utilizaram a ampla cobertura de radares meteorológicos nos Estados Unidos, juntamente com variáveis climáticas e topográficas, para estimar a precipitação numa grade fina de 4 km.

1.2 Contribuições propostas neste trabalho

Esta tese apresenta os softwares EDDA e EPPA, bem como métodos que foram propostos no seu desenvolvimento.

Embora já existam no Brasil e no exterior ferramentas computacionais para visualização de descargas no âmbito meteorológico, o software EDDA está sendo avaliado operacionalmente no CEMADEN desde o final de 2012 possibilitando o monitoramento de atividade convectiva graças à geração de campos de densidade de ocorrência de descargas NS.

Uma nova metodologia de agrupamento espaço-temporal de descargas foi também desenvolvida para ser incorporada a uma futura versão do software EDDA. Esse agrupamento utiliza uma janela deslizante temporal, também aqui proposta. Uma nova versão do software EDDA, que possibilita estimação da chuva convectiva a partir de dados de descargas, está em vias de começar a ser avaliada operacionalmente, embora a função de estimação não tenha sido proposta no escopo deste trabalho.

No tocante à previsão de chuva, o software EPPA, que se encontra em implementação, representa uma abordagem inédita que utiliza previsões de modelos numéricos. Esse modelo baseia-se num classificador (árvore de decisão) que foi treinado

com dados de radares meteorológicos, embora pretenda-se futuramente estender esse treinamento com dados de descargas NS.

2 DADOS METEOROLÓGICOS

O grande volume de dados meteorológicos disponíveis tem características como a heterogeneidade, diferentes resoluções espaciais e temporais e diferentes formatos. Essas características dificultam o estabelecimento de uma base de dados consistente, necessária à mineração de dados. Este trabalho utiliza dados de radar meteorológico com o objetivo de estimar as chuvas características de eventos convectivos. Assume-se aqui que descargas elétricas atmosféricas possam ser correlacionadas com atividade convectiva. Conseqüentemente, a atividade convectiva poderá ser monitorada e rastreada por meio de dados de descargas, os quais estão disponíveis para uma grande área do território brasileiro com uma resolução de milissegundos. Também são utilizados dados de modelos numéricos de previsão do tempo, sendo estes relativos a variáveis e índices de instabilidade selecionados. Este trabalho procura definir regras utilizando estas variáveis e índices de forma a fazer uma previsão de curto prazo da ocorrência de eventos convectivos.

2.1 Dados de descargas elétricas atmosféricas

Um raio (*flash*) pode ser composto de uma ou mais descargas elétricas atmosféricas (*strokes*). Raios estão tipicamente associados a tempestades. Segundo [Pinto e Pinto \(2008\)](#), não se conhece exatamente como as nuvens de tempestade se tornam carregadas. Em parte, isto se deve ao fato de que a estrutura elétrica de uma nuvem de tempestade é bastante complexa, sendo o resultado de processos microfísicos, que atuam em escalas de quilômetros, e processos microfísicos, que atuam em escalas de milímetros, ambos ocorrendo simultaneamente dentro da nuvem. Como resultado destes processos, cargas intensas são produzidas no interior da nuvem, que podem dar origem aos raios.

No processo indutivo, o campo elétrico atua na separação de cargas, através da polarização das partículas de gelo maiores como o granizo. A colisão destas partículas com as partículas de gelo menores, como os cristais de gelo, transfere cargas do granizo para os cristais. Para um campo elétrico orientado em direção descendente na atmosfera, o granizo transferirá cargas positivas para os cristais de gelo, tornando os cristais carregados positivamente e ficando carregado negativamente. Já o processo termoelétrico estabelece que a polaridade da carga transferida durante uma colisão entre diferentes partículas de gelo depende da temperatura no local da colisão. A [Figura 2.1](#) ilustra a estrutura de uma nuvem com seus centros de carga.

Um raio pode ser classificado de acordo com sua origem e o destino como sendo

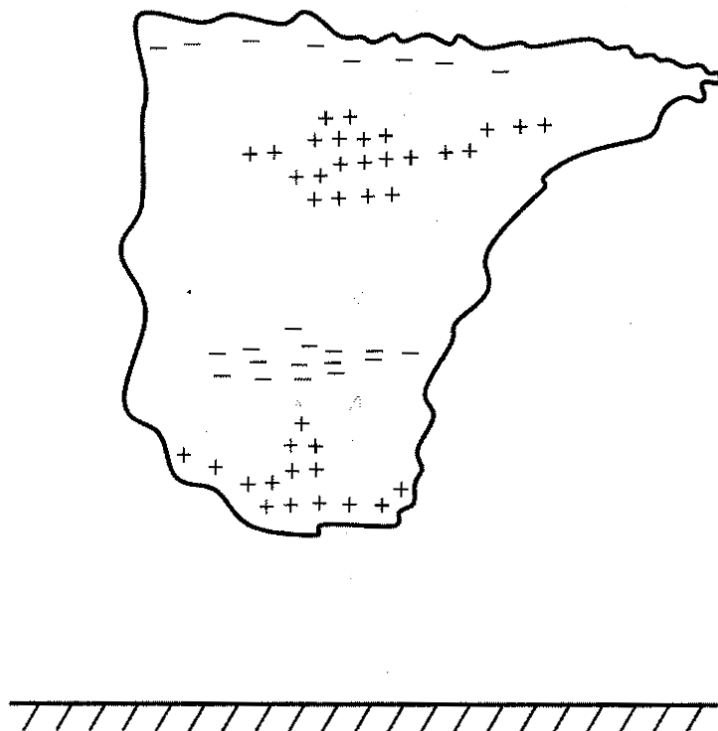


Figura 2.1 - Estrutura elétrica de uma nuvem de tempestade
 Fonte: Pinto e Pinto (2008)

nuvem-solo (NS), solo-nuvem, intra-nuvem (IN), nuvem-nuvem (NN) e nuvem-ar. Em função da utilização de dados do RINDAT, rede projetada para a detecção de descargas tipo NS e NN, este trabalho considera apenas descargas NS. Isto parece ser adequado para o estudo da atividade convectiva (LANG; RUTLEDGE, 2011).

Os raios NS podem também ser classificados como negativos ou positivos, dependendo do sinal da carga que é transferida da nuvem para o solo. Um raio NS inicia-se pela ocorrência de descargas fracas que ocorrem dentro da nuvem, quebrando a rigidez dielétrica do ar (tornando-o condutor ao invés de isolante). Em seguida, uma descarga fraca, denominada líder escalonado, se propaga da nuvem para o solo, buscando o melhor caminho, podendo se ramificar. Uma vez perto do solo, uma descarga conectante sai do solo em direção ao líder escalonado, geralmente a partir de objetos pontiagudos como árvores, edifícios ou mesmo pessoas. No instante do encontro, começa a fluir uma corrente intensa denominada descarga de retorno. É essa corrente que é percebida pela sua emissão luminosa e em ondas de rádio. Essa corrente se propaga em direção à nuvem e ilumina os ramos anteriormente criados pelo líder

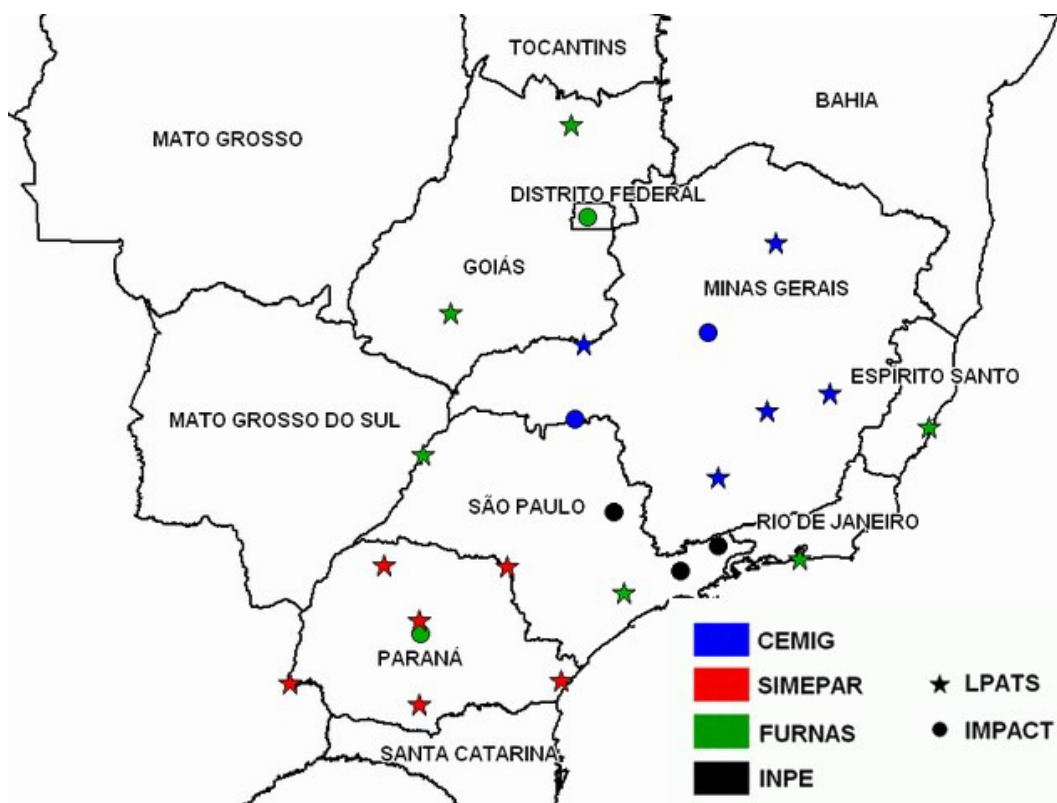


Figura 2.2 - Localização dos sensores da rede RINDAT
 Fonte: RINDAT (2013)

escalonado. Os valores típicos de corrente são entre 20 e 30 mil ampères, com uma duração de $100\mu s$. Em 80% dos casos, após a primeira descarga de retorno, outras descargas podem ocorrer, aproveitando o canal ionizado pelo líder escalonado.

Neste trabalho, são empregados dados de descargas ocorridas no estado de São Paulo no mês de janeiro de 2010 para análise de duas tempestades específicas que foram selecionadas como estudo de caso. Esses dados foram adquiridos pela rede RINDAT – Rede Integrada Nacional de Detecção de Descargas Atmosféricas (NACCARATO; PINTO, 2009), composta por uma rede de sensores de descargas distribuída no território nacional (Figura 2.2). O Grupo de Eletricidade Atmosférica do INPE (ELAT) fornece dados de descargas para o CPTEC/INPE (Centro de Previsão do Tempo e Estudos Climáticos) e para o CEMADEN (Centro de Monitoramento e Alerta de Desastres Naturais).

De acordo com Naccarato e Pinto (2009), a eficiência relativa de detecção de descargas NS é de cerca de 90% para o estado de São Paulo, considerando a rede de detectores RINDAT. Esta eficiência basicamente expressa uma probabilidade de detecção de

descargas para a rede.

Esses dados utilizam o formato UALF (*Universal ASCII Lightning Format*), que consiste em arquivos texto formatados em colunas de atributos tais como latitude, longitude e instante de ocorrência, polaridade e tipo (NN, NS e IN), sendo cada linha correspondente a uma descarga particular. Além de calcular a densidade de ocorrências, o software EDDA possibilita o cálculo da densidade de carga elétrica, uma vez que a carga q de cada descarga pode ser calculada a partir dos dados em formato UALF, conforme se segue:

$$q = \int I dt \sim \frac{1}{2} I_{\text{pico}} (t_{\text{subida}} + t_{\text{descida}}) \quad (2.1)$$

onde I_{pico} é a corrente de pico, e t_{subida} e t_{descida} são respectivamente os tempos de subida e descida da corrente. Sabe-se que a função $I = f(T)$ é aproximadamente linear, tanto na subida como na descida.

Em geral, os sensores de raios detectam componentes da onda eletromagnética esférica (EM) emitida por uma descarga. É necessária uma rede integrada de sensores para obter uma localização precisa. O sensor *Lightning Position and Tracking System* (LPATS) mede o tempo de chegada da onda EM, que é gravado de acordo com um referencial de tempo GPS (*Global Positioning System*), permitindo calcular a posição da descarga por triangulação. Por outro lado, o sensor *Improved Performance from Combined Technology* (IMPACT) também leva em conta a direção da descarga encontrada pelo método *Magnetic Direction Finding* (MDF), que detecta as duas componentes ortogonais do campo magnético (CUMMINS; MURPHY, 2009).

A partir dos dados de descargas individuais, para um dado período de tempo, e para uma determinada área, é possível gerar um campo de densidade de ocorrência de descargas, o qual pode ser visualizado conforme ilustrado na Figura 2.3.

2.2 Dados de radar meteorológico

Radares meteorológicos enviam pulsos eletromagnéticos e medem o sinal refletido por obstáculos como água e granizo. Quanto maior a gota e maior seu número, maior a parcela de energia retornada, o que permite estimar a quantidade de gotas de chuva em suspensão. O atraso entre a transmissão do pulso e seu eco permite estimar a distância. Alguns fatores podem gerar falsos ecos, como prédios (em baixas elevações), pássaros e insetos. Tempestades intensas também podem bloquear o sinal do

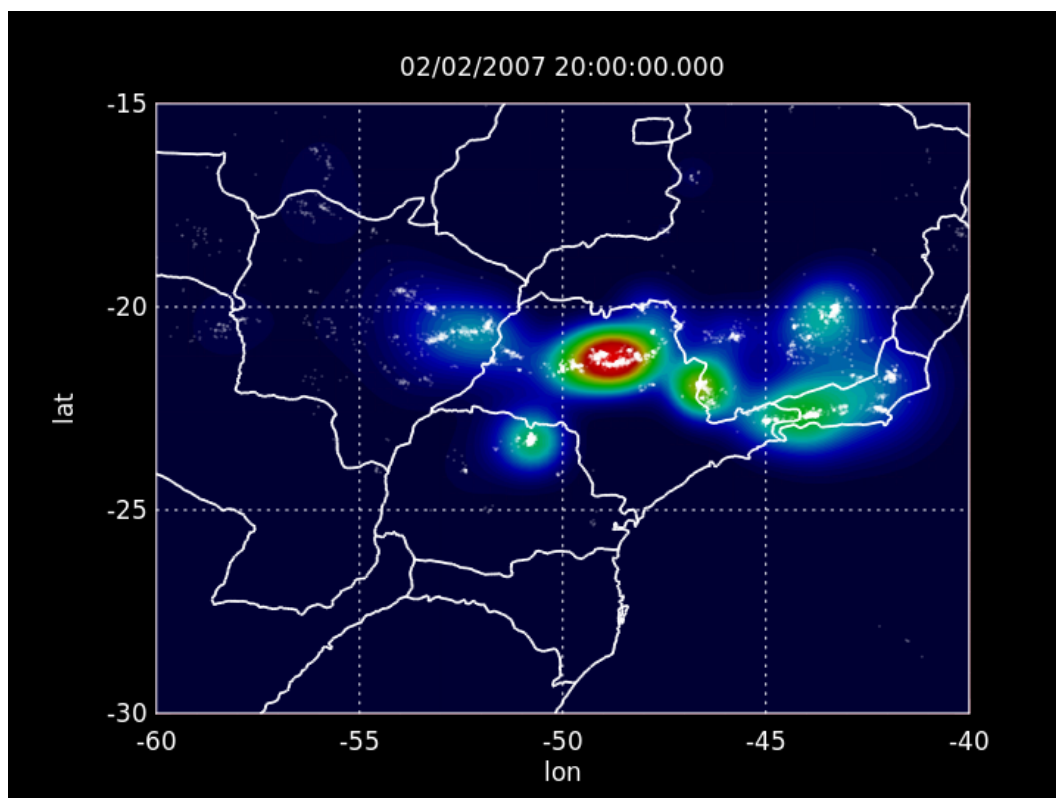


Figura 2.3 - Exemplo de visualização de campo de densidade de ocorrência de descargas elétricas atmosféricas, gerado pelo software EDDA.

radar, escondendo uma eventual chuva que pode estar acontecendo a uma distância maior. Por outro lado, nas regiões logo abaixo da linha de degelo, o derretimento de agregados (graupel e cristais de gelo) em zonas estratiformes acaba formando a chamada “banda brilhante”. À medida que um pulso de radar se distancia de sua origem, sua altura em relação a superfície da terra aumenta, chegando a atingir 12 km de altura a uma distância de 450 km. Dessa forma, distâncias acima de 450 km de observação estarão prejudicadas, pois o pulso eletromagnético estará muito acima das áreas de precipitação. Esses problemas podem ser minimizados a partir da fusão de imagem de radares próximos, como os localizados em Bauru e Presidente Prudente (Figura 2.4). Dessa maneira, um radar pode cobrir a deficiência do outro na região de sobreposição.

A taxa de precipitação pode ser estimada usando uma relação do tipo $Z - R$ apropriada (CALHEIROS; GOMES, 2010), expressa por $Z = 32 R^{1,65}$, onde Z é a refletividade do radar em dB e R a taxa de precipitação em mm/h. Esse tipo de relação varia de radar para radar, e de região para região, podendo ser obtida usando dados de precipitação verdadeira medida por pluviômetro ou de disdômetros.

Os dados de radar utilizados neste trabalho foram medidos por três radares do tipo Doppler que emitem pulsos eletromagnéticos na banda S. Essa banda é definida por frequências entre 2 e 4 GHz e engloba a chamada banda de 10 cm, que corresponde a comprimentos de onda próximos. Estes comprimentos de onda são da ordem de 10 vezes ou mais o diâmetro das gotas de chuva ou partículas de gelo, causando retro espalhamento do tipo Rayleigh, detectável pelo radar. Esses radares geram dados volumétricos, ou seja, possibilitam a localização em três dimensões de alvos, no caso, estruturas de precipitação. Além disso, o efeito Doppler permite medir a velocidade radial dos alvos. Dois destes radares pertencem ao IPMet/UNESP, sendo localizados nas cidades de Bauru e Presidente Prudente, alinhados na direção leste-oeste e localizados a 240 km de distância. Estes radares têm uma resolução radial de 250 m e de 1° em azimute, com um alcance útil de 240 km e realizam uma varredura a cada 7,5 min (HELD et al., 2010). O terceiro radar está localizado em São Roque, tem um alcance de 400 km, resolução radial de 125 m e 2° em azimute. Os dados dos três radares foram limitados inicialmente a 240 km, pois isso permite que as áreas de cobertura de ambos os radares do IPMet terminem sobre a cidade congênere. No entanto, para efeito dos cálculos de correlação, foram considerados apenas dados até uma distância máxima de 150 km de cada radar, uma vez que a atenuação do feixe do radar acima desta distância pode impedir a validação proposta.

Neste trabalho, pretendia-se usar dados correspondentes a vários verões de anos recentes, mas devido a dificuldades na obtenção de dados de radar meteorológico, consideraram-se inicialmente apenas os dados de janeiro/2010 para o verão de 2009/2010 e posteriormente dados de dezembro/2010 e janeiro/2011 e fevereiro/2011, correspondentes ao verão de 2010/2011. Para este verão, consideraram-se apenas os dados dos radares de Bauru e Presidente Prudente.

As imagens de radares meteorológicos empregadas neste trabalho correspondem a um corte horizontal numa dada altitude (3 km), *Constant Altitude Plan Position Indicator* (CAPPI). Essas imagens são obtidas por sucessivas varreduras do radar em diferentes elevações fixas, no chamado modo PPI (*Plan Position Indicator*). Cada varredura corresponde a um ângulo de elevação fixo, variando-se o ângulo azimutal de 0° a 360° . Em cada uma das imagens PPI resultantes, a altitude é função da distância ao radar, devido à inclinação do feixe do radar e os alvos observados são delimitados por dois círculos concêntricos com centro no radar. A composição das imagens PPI para a altitude desejada resulta na correspondente imagem CAPPI (MEISCHNER, 2003).

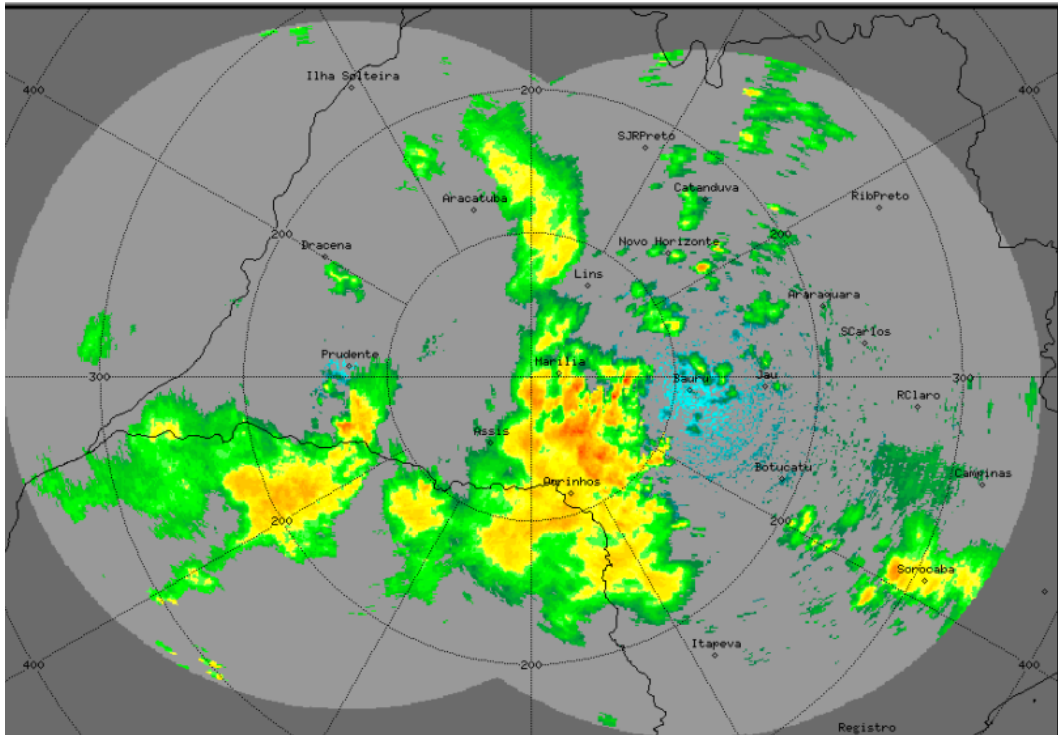


Figura 2.4 - Exemplo de imagem CAPPI composto dados dos radares meteorológicos de Bauru e Presidente Prudente (os círculos têm 240 km de raio)

No caso, foram utilizadas imagens CAPPI para uma altitude de 3 km. Esta altitude pode ser utilizada para classificar a precipitação convectiva ou estratiforme (STEINER et al., 1995). Outros trabalhos têm empregado outras altitudes, como o de Mosier et al. (2011) focalizado em altitude de formação de gelo. Os dados CAPPI dos três radares foram interpolados para uma grade regular de $0,02^\circ$ e uma resolução temporal de 10 min. Esta resolução temporal foi escolhida para ser um múltiplo da taxa de atualização de 5 min dos dados de raios.

Pixels correspondentes à precipitação convectiva foram identificados utilizando os critérios de classificação específica de Steiner et al. (1995), que se baseia em valores de refletividade e diferença do pixel com seu entorno. Um algoritmo região padrão crescente foi então aplicada a esses pixels para identificar e localizar as “estruturas de precipitação”, que correspondem às células de precipitação convectiva.

Neste trabalho, por simplicidade, a evolução temporal das estruturas convectivas foi rastreada verificando-se estruturas sobrepostas em imagens sucessivas. De acordo com Lakshmanan e Smith (2010), este rastreamento com base na sobreposição produz bons resultados em termos de consistência da evolução temporal da estrutura associada à variável observada e também fornece trajetórias quasi-lineares. No en-

tanto, o rastreamento baseado em sobreposição pode erroneamente fracionar uma trajetória em duas trajetórias separadas mais curtas. Uma comparação de esquemas de rastreamento estaria fora do escopo deste trabalho pois exigiria uma extensa base de dados de trajetórias. A título de exemplo, um sistema bem conhecido em Meteorologia, o TITAN – *Thunderstorm Identification, Tracking, Analysis and Nowcasting* (DIXON; WIENER, 1993), correlaciona centroides em imagens sucessivas.

2.3 Dados de modelos numéricos

Modelos de previsão numérica de tempo simulam a evolução de variáveis atmosféricas numa grade a partir de equações físicas básicas que regem seu comportamento. Na fase de assimilação, dados provenientes de medições meteorológicas são combinados ao resultado da última previsão, num processo chamado de “assimilação de dados”. A assimilação fornece a condição inicial do modelo, denominada “análise”. Mesmo levando em conta as incertezas envolvidas, os dados de análise são uma importante ferramenta, por estimarem o estado da atmosfera numa grade regular mesmo na ausência de dados coletados. A partir da análise, as equações físicas são integradas no espaço e no tempo, gerando as previsões sucessivas. Este trabalho utiliza dados do modelo numérico ETA.

O modelo de mesoescala de previsão de tempo ETA foi originalmente desenvolvido pela universidade de Belgrado em conjunto com o Instituto de Hidrometeorologia da Iugoslávia (MESINGER et al., 1988; BLACK, 1994). O modelo ETA vem sendo utilizado operacionalmente pelo CPTEC/INPE desde 1996. Algumas variáveis prognósticas do modelo são: temperatura virtual do ar, componente zonal e meridional do vento, umidade específica, pressão à superfície e energia cinética turbulenta.

Dentre essas variáveis prognósticas, devem-se selecionar as que serão usadas como atributos na análise. A modelagem de mecanismos de formação de tempestades pode fornecer alguns indicadores, assim como a análise de registros de eventos convectivos severos ocorridos no passado. Especialistas em previsão também podem ser consultados para a escolha dessas variáveis e índices. No caso deste trabalho, essa escolha foi feita de acordo com a Ferramenta Objetiva de Previsão do Tempo do CPTEC/INPE (ANDRADE et al., 2010). Dessa maneira, as mesmas variáveis e índices utilizados naquela ferramenta foram utilizadas neste trabalho. As variáveis extraídas das saídas do modelo foram:

- $umrl_{500}$, $umrl_{700}$, $umrl_{850}$, $umrl_{925}$, $umrl_{1000}$: Umidade relativa nos níveis de 500 hPa, 700 hPa, 850 hPa, 925 hPa e 1000 hPa, respectivamente;

- $t_{500}, t_{700}, t_{850}$: Temperatura nos níveis de 500 hPa, 700 hPa e 850 hPa, respectivamente;
- $uvel_{500}, uvel_{850}$: Velocidade zonal do vento nos níveis de 500 hPa e 850 hPa, respectivamente;
- $vvel_{500}, vvel_{850}$: Velocidade meridional do vento nos níveis de 500 hPa e 850 hPa, respectivamente.

Empregam-se aqui umidades relativas médias, expressas pelas equações 2.2 e 2.3, e índices de instabilidade derivados das variáveis acima descritas, sendo que Td denota a temperatura de ponto de orvalho. Os índices de instabilidade permitem inferir o grau de instabilidade da atmosfera, que pode ser absolutamente estável, condicionalmente instável ou absolutamente instável. Os índices VT, CT, TT e K estão associados ao gradiente vertical de temperatura, sendo descritos pelas equações 2.4, 2.5, 2.6 e 2.7, respectivamente.

O índice VT, proposto por Miller (1972), é dado pela diferença das temperaturas nos níveis de 500 e 850 hPa. O índice CT é dado pela diferença entre a temperatura do ponto de orvalho no nível de 850 hPa (td_{850}), a qual representa a umidade neste nível, e a temperatura no nível de 500 hPa (t_{500}). Pelo fato da magnitude de Td_{850} ser maior do que a de t_{500} , este índice é mais fortemente influenciado pelo primeiro termo. O índice TT nada mais é que a soma dos índices CT e VT.

O índice K, proposto por George (1960), pode ser interpretado como uma medida do potencial de ocorrência de tempestade, sendo baseado na taxa vertical de variação de temperatura, no conteúdo de umidade na baixa troposfera e na extensão vertical da camada úmida. Este índice é muito usado para avaliar chuvas fortes, pois a presença de camadas úmidas em 850 e 700 hPa está associada à grande quantidade de água precipitável.

Seguem-se as expressões das variáveis e índices de instabilidade citados:

- Umidade relativa média dos níveis entre 500 e 850 hPa:

$$umrl_{500-850} = \frac{umrl_{500} + umrl_{700} + umrl_{850}}{3} \quad (2.2)$$

- Umidade relativa média dos níveis entre 500 e 1000 hPa:

$$umrl_{850-1000} = \frac{umrl_{850} + umrl_{925} + umrl_{1000}}{3} \quad (2.3)$$

- Vertical Totals (VT):

$$VT = t_{850} - t_{500} \quad (2.4)$$

- Cross Totals (CT):

$$CT = td_{850} - t_{500} \quad (2.5)$$

- Total Totals (TT):

$$TT = VT + CT \quad (2.6)$$

- K Index (K):

$$K = (t_{850} - t_{500}) + td_{850} - (t_{700} - td_{700}) \quad (2.7)$$

Além das variáveis e índices acima descritos, este trabalho usa a variável ω_{500} e os índices SWEAT, CAPE, CINE e BLI, descritos a seguir.

O índice SWEAT, *Severe Weather Threat Index* (índice de ameaça de tempo severo), também foi desenvolvido por Miller (1972), sendo muito utilizado para analisar o potencial de tempestades muito severas (formações de tornados). Ele é mais refinado que o TT, pois combina a termodinâmica da atmosfera com a dinâmica, ao levar em conta o cisalhamento horizontal da atmosfera. Valores altos dos índices TT, K e SWEAT estão associados à maior probabilidade de ocorrência de tempestades. Este índice é calculado como:

$$\begin{aligned} SWEAT = 12 \times td_{850} + 20 \times TERM_2 + \\ + 2 \times SKT_{850} + SKT_{500} + SHEAR \end{aligned} \quad (2.8)$$

onde,

- $TERM_2 = \max(TT - 49, 0)$
- SKT_{500}, SKT_{850} : Velocidade do vento em nós nos níveis de 500 e 850 hPa, respectivamente;
- $SHEAR = 125 \times [\text{seno}(DIR_{500} - DIR_{850}) + 0, 2]$
- DIR_{500}, DIR_{850} : Direção do vento em graus nos níveis de 500 e 850 hPa, respectivamente.

Valores de SWEAT acima de 300 indicam tempestades severas e, acima de 400, tornados. Entretanto, para que o SWEAT seja válido, os seguintes critérios devem ser obedecidos:

- Direção do vento no nível de 850 hPa no intervalo de 130 a 250° (SE a SW no Hemisfério Norte) e 310 e 70 (NW a NE no Hemisfério Sul);
- Direção do vento no nível de 500 hPa no intervalo de 210 a 310° (W-NW);
- Diferença entre a direção dos ventos positiva (Advecção Quente);
- Velocidades dos ventos nos níveis de 500 ou 850 hPa maiores que 15 nós (7,72 m/s).
- Todos os termos da equação 2.8 positivos.

A variável ω_{500} e os demais índices de instabilidade são calculados diretamente pelo modelo, ou seja, aparecem diretamente nas saídas do modelo.

- ω_{500} : Velocidade vertical do vento no nível de 500 hPa.
- *CAPE* (*Convective Available Potential Energy*): este índice, como o próprio nome indica, mede a energia potencial convectiva disponível numa parcela de ar em equilíbrio estático, estando relacionado à diferença entre as temperaturas da parcela e a temperatura da atmosfera à sua volta, sendo que quanto maior esta diferença, maior o empuxo, que pode forçar o ar úmido a subir ocasionando a condensação do vapor d'água e a consequente liberação de calor latente.

$$CAPE = \int_{NCC}^{NEN} g \frac{tp - ta}{ta} dz \quad (2.9)$$

onde:

NEN é o Nível de Empuxo Neutro (próximo à tropopausa);

NCC é o Nível de Condensação Convectiva;

tp é a temperatura da parcela;

ta é a temperatura do ambiente;

g é a aceleração da gravidade.

- *CINE (Convective Inhibition Energy)*: este índice, a energia de inibição convectiva, é análogo ao índice CAPE, porém mede a energia que a parcela de ar considerada tem que vencer desde a superfície até o NCC para dar início à convecção.

$$CINE = \int_{SUP}^{NCC} g \frac{tp - ta}{ta} dz \quad (2.10)$$

onde SUP é o nível de SUPERfície.

- *BLI (Best Lifted Index)*: o índice de levantamento melhorado é um aperfeiçoamento do Lifted Index (LI), índice proposto por Galway (1956), sendo expresso por:

$$LI = t_{500} - tp_{500} \quad (2.11)$$

onde tp_{500} é a temperatura da parcela de ar em 500 hPa.

O LI é idêntico ao índice Showalter (SWI), exceto pelo fato de que, considerando-se o diagrama termodinâmico SKEW-T, a parcela de ar é “levantada” pela adiabática seca a partir de um determinado nível, até que esta cruze a linha de razão de mistura saturada que parte de td , sendo que deste ponto ela “segue” pela adiabática úmida até o nível de 500 hPa. O índice LI é muito útil para indicar tempestades severas (um valor menor ou igual a -6 é considerado crítico). Calculam-se diversos valores do índice LI correspondentes a diversos níveis aleatórios a partir dos quais a parcela de ar é “levantada”. O valor do índice BLI é dado pelo “melhor” destes valores de LI, ou seja, o menor valor, que corresponde à condição mais instável.

Por fim, o cálculo dos is índices de instabilidade pressupõe que as seguintes hipóteses sejam atendidas:

- a) A parcela de ar considerada é considerada um sistema fechado sem troca de massa com a atmosfera à sua volta;
- b) A parcela de ar considerada é termicamente isolada, ou seja, pode ascender ou descender sem trocar calor com a atmosfera à sua volta, sofrendo apenas uma evolução adiabática;

- c) A pressão interna da parcela de ar considerada é igual à da atmosfera ao seu redor;
- d) A atmosfera está em equilíbrio hidrostático, mas não necessariamente a parcela de ar considerada.
- e) A velocidade vertical da parcela de ar considerada é baixa, de forma que a correspondente energia cinética não é significativa no balanço total de sua energia interna.

Os dados de modelo abrangem os mesmos períodos considerados para os dados de radar meteorológico. Assim, neste trabalho, foram utilizados dados do modelo ETA 20 km (resolução 0,2°) para o período de 1 a 26 de janeiro de 2010. Para cada dia considerado, dispõe-se da análise da 00 UTC, e previsões a cada 6 horas até 90 horas de previsão. A faixa de latitudes disponível é de -35° a -15° e a faixa de longitudes, de -60° a -40° . O número de pontos de grade disponível é de 101 x 101.

Também foram utilizados dados do modelo ETA 5 km Serra do Mar (resolução 0,05°) para o período de 1 a 26 de janeiro de 2010, assim como para os meses de dezembro/2010 e janeiro/2011 e fevereiro/2011. Para cada dia considerado dispõe-se da análise das 00 UTC, e previsões a cada hora até 47 horas de previsão. A faixa de latitudes disponível é de -27° a $-18,85^\circ$ e a faixa de longitudes, de -53° a $-39,55^\circ$. O número de pontos de grade disponíveis é de 270 x 164. No restante deste texto, este modelo será referido simplesmente como modelo ETA 5 km.

Para as análises sinóticas das tempestades estudadas, utilizou-se o Modelo de Circulação Geral Atmosférico (AGCM) do CPTEC (BONATTI, 1996). Este modelo tem origem no modelo MRF (Medium Range Forecasting Model) desenvolvido pelo NCEP (National Centers for Environmental Prediction). O AGCM tem sido utilizado operacionalmente no CPTEC tanto para previsão de tempo quanto no escopo de previsão de clima (CAVALCANTI, 1996).

3 ALGUNS MÉTODOS RELACIONADOS A ESTE TRABALHO

Abordam-se aqui alguns métodos que foram usadas neste trabalho. O software EDDA, que gera campos de densidade de descargas, utiliza o método de estimação de densidade por kernel gaussiano descrito na Seção 3.1.

A Seção 3.2 aborda métodos de agrupamento e classificação de dados. O método de agrupamento DENCLUE, que foi efetivamente utilizado no software EDDA, é descrito na Seção 3.2.1.

A Seção 3.2.2 descreve o método de classificação utilizado no software EPPA, as árvores de decisão. Esse software foi inspirado na Ferramenta Objetiva de Previsão do Tempo do CPTEC/INPE, descrita na Seção 3.3.

A seguir, a Seção 3.4 apresenta uma análise da correlação entre o número de descargas NS e a chuva convectiva. A Seção 3.5 mostra a estimação de precipitação, assim como a estimação da precipitação acumulada a partir do número de descargas utilizando o software EDDA.

3.1 Estimação de densidade

A estimação de densidade permite gerar um campo de densidade de ocorrência a partir de um conjunto de eventos amostrados. No caso deste trabalho, tais eventos são as descargas, sendo estimado um campo fictício de densidade de ocorrência de descargas para o intervalo de tempo considerado.

O estimador de kernel bidimensional clássico (simétrico em relação às duas dimensões) pode ser escrito como (SILVERMAN, 1986):

$$\hat{f}(X) = \frac{1}{nh^2} \sum_{i=1}^n K\left(\frac{d(X, X_i)}{h}\right) \quad (3.1)$$

onde:

- X é o par de coordenadas (x, y) para uma grade bidimensional,
- $\hat{f}(X)$ é a estimativa de densidade no ponto de grade X ,
- n é o número de amostras,
- $d(X, X_i)$ é a distância euclidiana do ponto de grade X ao evento X_i ,

- $K(r)$ é a função de kernel bidimensional, onde r representa a distância euclidiana normalizada pelo parâmetro h ,
- h é a largura janela de suavização centrada no ponto de grade X , também chamada raio de influência.

No caso do kernel gaussiano bidimensional e simétrico:

$$\hat{f}(X) = \frac{1}{2\pi nh^2} \sum_{i=1}^n \exp \left[\frac{(x - x_i)^2 + (y - y_i)^2}{2h^2} \right] \quad (3.2)$$

O parâmetro h controla a suavidade do campo e pode ser escolhido de forma manual, caso a caso, de maneira a encontrar um resultado visual satisfatório. Quanto maior seu valor, mais suave será o campo gerado. Entretanto, um esquema automático para ajuste de h foi proposto em [Silverman \(1986\)](#), no qual o valor de h é calculado de forma a minimizar o valor médio do erro quadrático integrado (*Mean Integrated Squared Error* – MISE):

$$\text{MISE}(\hat{f}) = E \int \{\hat{f}(X) - f(X)\}^2 dx \quad (3.3)$$

Assumindo-se uma distribuição gaussiana, o valor ótimo de h que minimiza o MISE é expresso em função do desvio padrão σ da gaussiana como sendo:

$$h_{\text{opt}} = n^{-1/6} \sigma \quad (3.4)$$

Uma possível escolha para σ é a raiz quadrada da média das variâncias das coordenadas dos eventos amostrados em cada dimensão:

$$\sigma^2 = \frac{1}{2}(\sigma_x^2 + \sigma_y^2) \quad (3.5)$$

onde:

$$\sigma_x^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \left(\frac{1}{n} \sum_{i=1}^n x_i \right)^2$$

$$\sigma_y^2 = \frac{1}{n} \sum_{i=1}^n y_i^2 - \left(\frac{1}{n} \sum_{i=1}^n y_i \right)^2$$

Um esquema análogo (SCOTT, 1992), considera o caso não-simétrico, calculando-se a largura de janela h e o desvio padrão σ separadamente para cada dimensão (nesse caso a função kernel é unidimensional):

$$\hat{f}(X) = \frac{1}{nh_x h_y} \sum_{i=1}^n K_1\left(\frac{x - x_i}{h_x}\right) K_1\left(\frac{y - y_i}{h_y}\right) \quad (3.6)$$

onde:

$$K_1(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$$

$$h_x = n^{-1/6} \sigma_x$$

$$h_y = n^{-1/6} \sigma_y$$

Os esquemas acima descritos calculam um valor constante da largura de janela h para todos os pontos de grade. (SILVERMAN, 1986) propõe que melhores resultados podem ser obtidos variando o valor de h conforme a densidade no ponto de grade considerado:

$$\hat{f}(X) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h_i^2} K\left(\frac{d(X, X_i)}{h_i}\right) \quad (3.7)$$

onde:

$$h_i = h \lambda_i$$

$$\lambda_i = (\tilde{f}(X_i)/g)^{-\alpha}$$

onde h é a janela ótima global, h_i é a largura de janela local, λ_i é o fator de largura de janela local, $\tilde{f}(x_i)$ é uma estimativa prévia da densidade local usando janela fixa e α é um parâmetro entre 0 e 1 que define o grau de influência da densidade local sobre a janela local. Os λ_i são normalizados por g , que é a média geométrica de $\tilde{f}(x_i)$

Outra possibilidade, proposta em Politi et al. (2006) é calcular uma largura de janela $h(X)$ para cada ponto de grade:

$$\hat{f}(X) = \frac{1}{nh(X)^2} \sum_{i=1}^n K\left(\frac{d(X, X_i)}{h(X)}\right) \quad (3.8)$$

onde:

$$h(X) = n^{-1/6} \sigma(X)$$

$$\sigma^2(X) = \frac{1}{n} \sum_{i=1}^n d(X, X_i)^2 - \left(\frac{1}{n} \sum_{i=1}^n d(X, X_i) \right)^2$$

3.1.1 Avaliação do uso de largura de janela fixa ou adaptativa

Os diversos esquemas para definir a largura de janela h acima expostos foram avaliados utilizando-se um campo de densidade gerado aleatoriamente. A partir deste campo, foi gerado um conjunto de pontos representando eventos amostrados, localizados em coordenadas geradas aleatoriamente, obedecendo à essa distribuição de densidade. Utilizando-se esse campo de densidade como referência, procurou-se reproduzi-lo a partir desse conjunto de eventos, utilizando esses diversos esquemas.

Para gerar um campo de densidade de eventos de maneira aleatória numa grade regular bidimensional, gerou-se um relevo que lembra uma cadeia de montanhas, utilizando um método fractal, correspondente ao algoritmo “subdivisão diamante-quadrado” (MILLER, 1986), empregado em computação gráfica. Assim, a altitude do relevo corresponde ao valor da densidade no ponto de grade considerado.

No algoritmo “subdivisão diamante-quadrado”, subdivide-se uma grade quadrada, calculando-se o valor do centro de cada quadrado. O resultado é uma grade de “diamantes” (Figura 3.1, esquerda). Novamente, calcula-se o valor do centro de cada diamante. O resultado é uma nova grade quadrada. (Figura 3.1, direita). Repete-se a operação até o nível de refinamento desejado. A alternância entre quadrados e diamantes evita efeitos sistemáticos indesejáveis de uma grade regular. Para calcular o valor do centro de cada diamante ou quadrado, toma-se a média dos vértices e adiciona-se um deslocamento aleatório. A amplitude desse deslocamento decresce exponencialmente à medida que a malha é refinada. Esse fator de redução controla a rugosidade (ou dimensão fractal) da superfície final. A Figura 3.2 (superior) mostra o relevo fractal gerado pelo método, enquanto que a Figura 3.2 (centro) mostra o correspondente campo de densidade na forma de curvas de nível e a Figura 3.2 (inferior) mostra os eventos individuais gerados.

Para gerar as coordenadas dos eventos de maneira aleatória, interpretou-se o campo de densidade de eventos gerado anteriormente como uma matriz de densidade de probabilidade em duas dimensões. Essa densidade é reduzida a uma única dimensão ao integrar-se a densidade de probabilidade no sentido horizontal, gerando a densidade de probabilidade vertical. Considerando-se a coordenada vertical (y) de um evento, a correspondente densidade de probabilidade horizontal (na dimensão x) é extraída de um corte horizontal da matriz de densidade original (2D).

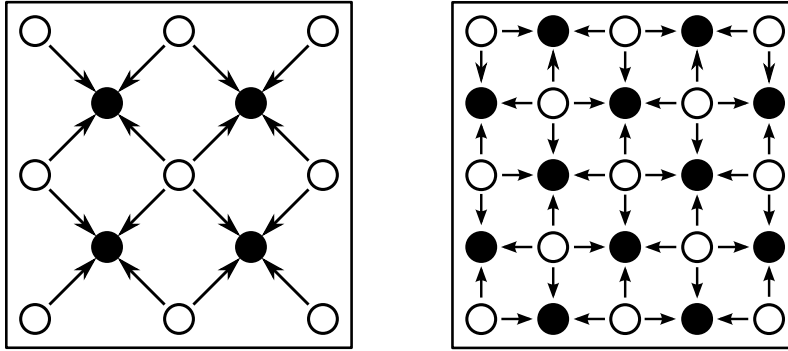


Figura 3.1 - Subdivisão de uma grade regular pelo algoritmo diamante-quadrado. O valor do campo nos círculos pretos é calculado em função dos valores dos círculos brancos adjacentes.

Tabela 3.1 - Erro médio quadrático integrado (MISE) dos vários esquemas de janela fixa e adaptativa para o conjunto de eventos considerado.

Esquema adotado	MISE x 10^{-5}
Janela fixa (Equação 3.5)	2,17
Janela fixa (Equação 3.6)	2,13
Janela adaptativa (Equação 3.7)	1,71
Janela adaptativa (Equação 3.8)	1,87

Para escolher de forma aleatória os eventos e suas correspondentes coordenadas a partir da densidade de probabilidade vertical, utilizou-se o método de [Devroye \(1986\)](#).

Uma vez obtidas a latitude e longitude das descargas, aplicou-se cada um dos métodos de largura de janela. Para cada um dos esquemas de largura de janela descritos pelas Equações 3.4 a 3.7, foram estimados campos de densidade a partir desse mesmo conjunto de eventos aleatórios e calculou-se para cada um o MISE (Equação 3.2) relativo ao campo de densidade original. Constatou-se que os melhores resultados foram obtidos com larguras de janela adaptativas, conforme observado na Figura 3.3 que apresenta as diversas estimativas do campo de densidade, e também na Tabela 3.1 que apresenta o MISE resultante para esses vários esquemas.

Esses resultados mostram a viabilidade do uso de estimação de densidade para eventos simulados análogos às descargas elétricas. Embora o desempenho de estimação de densidade de esquemas de largura de janela adaptativos seja teoricamente melhor, o restante deste trabalho utiliza larguras de janelas fixas, uma vez que foi constatado pelos meteorologistas que a escolha manual da largura de janela permite ajustar a

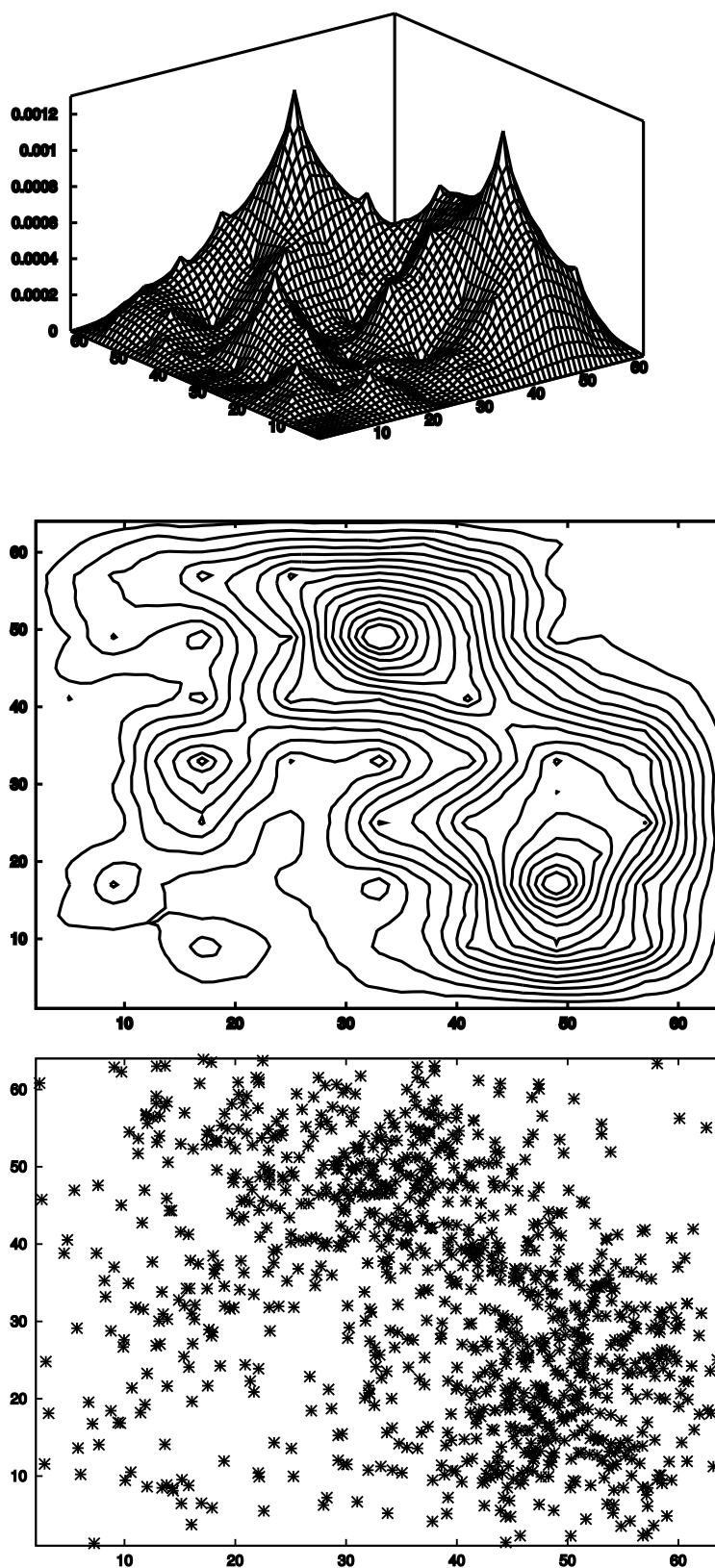


Figura 3.2 - O campo de densidade aleatório gerado método fractal para teste (acima), curvas de nível correspondente (meio) e conjunto aleatório de eventos gerado a partir desse campo (em baixo).

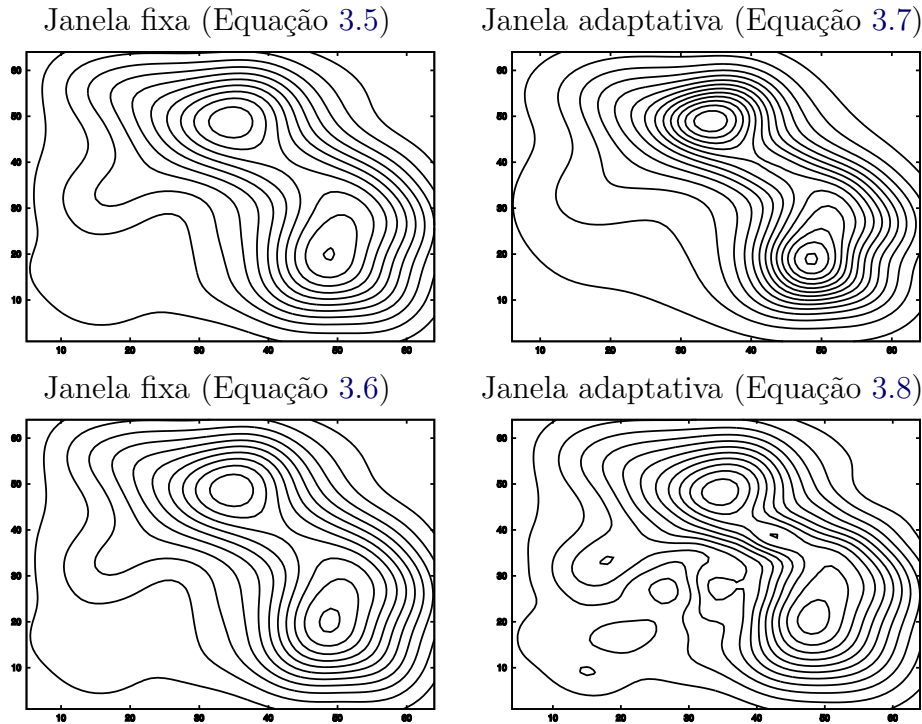


Figura 3.3 - Campos de densidade estimados usando os vários esquemas de ajuste de largura de janela a partir do conjunto aleatório de eventos considerado.

suavização do campo de densidade à escala do fenômeno convectivo analisado. Dessa forma, possibilitaria correlacionar melhor descargas e atividade convectiva.

3.2 Métodos de agrupamento e classificação

Este trabalho utiliza técnicas de aprendizado de máquina originadas na Inteligência Artificial (IA), especificamente, técnicas de aprendizado de máquina tais como agrupamento e classificação. Segundo [Bittencourt \(2013\)](#), algumas das principais áreas atuais da IA são: sistemas especialistas, visão computacional, robótica, controle inteligente, lógica nebulosa, inteligência artificial distribuída, linguagem natural, redes neurais artificiais e algoritmos genéticos.

Um aspecto da IA é a capacidade de aprendizado e adaptação, característica das técnicas de aprendizado de máquina. De maneira geral, métodos de agrupamento podem ser divididos em supervisionados ou não-supervisionados, conforme as classes sejam conhecidas *a priori* ou não, respectivamente. Pode-se então dizer que o agrupamento não-supervisionado é uma forma de aprendizado por observação, enquanto a classificação é uma forma de aprendizado por exemplos ([HAN; KAMBLER, 2011](#)).

A descoberta de conhecimento em bancos de dados (*Knowledge Discovery in Databases* – KDD) é um processo para extrair informações úteis (conhecimento) de grandes volumes de dados (FAYYAD et al., 1996). A Figura 3.4 ilustra as etapas deste processo. Inicialmente, é necessário que os dados estejam num formato acessível computacionalmente, possivelmente armazenados num Sistema de Gerenciamento de Banco de Dados Relacional (SGBD-R), ou mesmo na forma de tabelas ASCII ou matrizes binárias. Na fase de preparação dos dados, selecionam-se os atributos e as instâncias relevantes. Isso pode ser feito levando em conta um conhecimento prévio do domínio do problema, ou utilizando métodos automáticos. Na etapa de pré-processamento, realiza-se um controle de qualidade dos dados, de forma a tratar dados considerados como anomalias (*outliers*), dados incompletos e inconsistentes, pois é natural ocorrerem falhas nos processos de aquisição de dados reais, sejam eles manuais ou automáticos. Na etapa de transformação, convertem-se os dados brutos num formato padrão, adequado para o seu uso pelo método de mineração selecionado. Pode ser feita uma limiarização de certos atributos, convertendo valores contínuos em discretos (por exemplo, dividindo-se a intensidade da chuva em faixas correspondentes à chuva fraca, média e forte). Também pode ser feita a normalização, para evitar que um atributo tenha maior influência do que outros, no caso de métodos que sejam sensíveis a esse problema. Outra técnica é combinar dois ou mais atributos correlacionados num único atributo, reduzindo a dimensionalidade sem perder informação. Finalmente, a etapa de mineração de dados consiste em aplicar técnicas de sumarização, classificação, regressão, associação ou agrupamento aos dados. Os padrões ou relações descobertos são analisados e avaliados, gerando conhecimento. Cada uma das etapas descritas pode ser refeita e aperfeiçoada durante esse processo, na tentativa de se obter melhores resultados.

Nesse contexto, a etapa de mineração de dados pode ser definida como o processo de descobrir padrões em dados. Os padrões descobertos devem ser significativos e trazer eventualmente uma nova informação relativa aos dados. Nesta etapa, aplicam-se tipicamente técnicas de reconhecimento de padrões, estatísticas e de IA. O objetivo do processo KDD é descobrir conhecimento novo ou oculto em grandes bases de dados. O processo pode ser automático ou (mais usualmente) semi-automático (WITTEN; FRANK, 2000).

Nesse contexto, padrão pode ser definido como um conjunto de dados com valores específicos ou com determinadas características que aparecem frequentemente e que possa ser associado a um determinado evento. Neste trabalho, por exemplo, buscavam-se padrões formados por índices de instabilidade e variáveis atmosféricas que possam

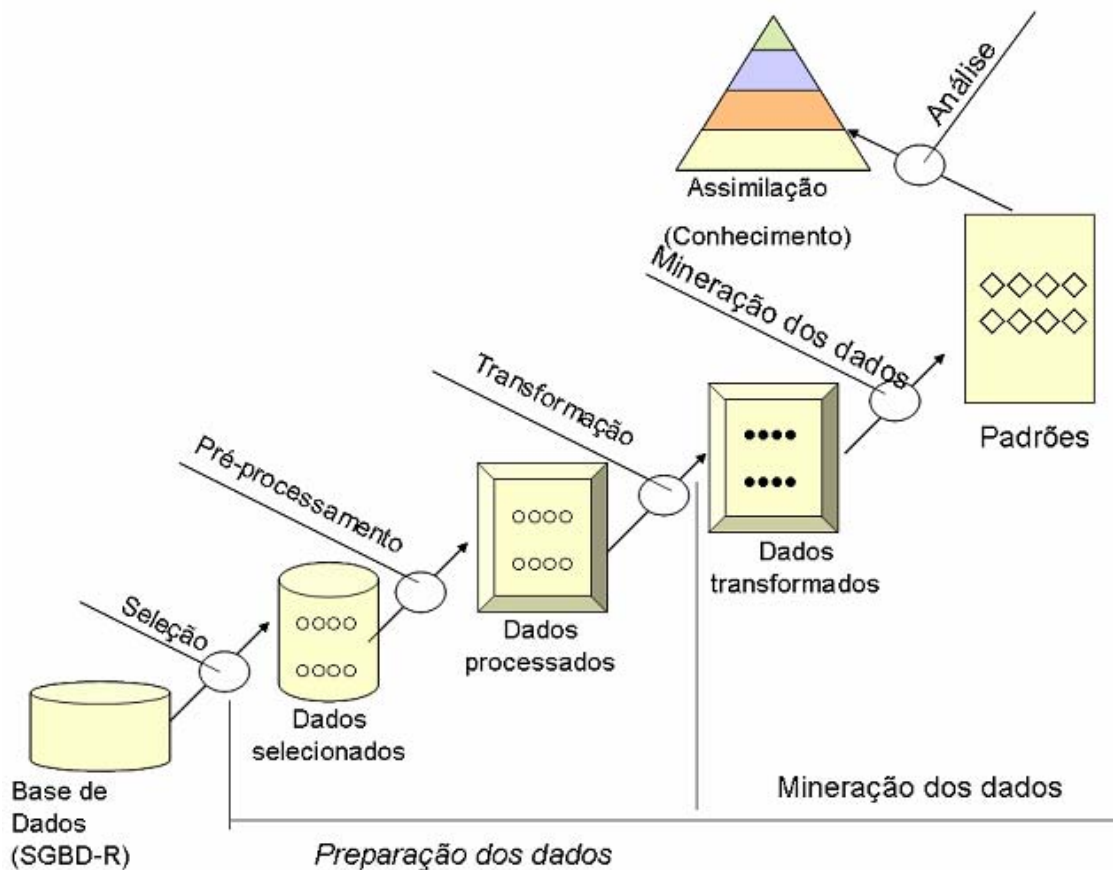


Figura 3.4 - Fluxo do processo de descoberta de conhecimento em banco de dados (KDD).
 Fonte: Adaptada de Fayyad et al. (1996)

ser associados à ocorrência de chuva forte e convectiva. Uma vez identificados estes padrões para casos conhecidos de chuva forte e convectiva, pode-se empregá-los na previsão de chuva por meio do monitoramento das saídas de um modelo de previsão de tempo.

3.2.1 O método de agrupamento DENCLUE

Agrupamento (*clustering*) é o processo de agrupar instâncias em grupos (*clusters*) ou classes (HAN; KAMBLER, 2011), de tal maneira que instâncias dentro de um mesmo grupo tenham alta similaridade entre si, mas sejam bem dissimilares a instâncias de outros grupos. O grau de similaridade é definido por alguma métrica que emprega os valores dos atributos selecionados para descrever as instâncias. Métodos de agrupamento encontram as mais diversas aplicações, incluindo pesquisa de mercado, reconhecimento de padrões, biologia, segurança, processamento de imagens e busca

de documentos na Web. No caso deste trabalho, as instâncias são as ocorrências de descargas elétricas atmosféricas e utilizaram-se limites de distância e de intervalo de tempo entre descargas como critério de agrupamento.

Dentre os vários métodos de agrupamento existentes, foi selecionado o método DENCLUE (HINNEBURG; GABRIEL, 2007), pois este método baseia-se na geração de campos de densidade de instâncias gerado por estimação de *kernel* (Seção 3.2.1). Nesse método de agrupamento, é feita uma correspondência entre cada grupo e um dos máximos locais da densidade de instâncias estimada. Cada instância é o ponto de partida de uma busca pelo método de subida de encosta (*hill climbing*), que converge para o máximo local de densidade de instâncias do grupo ao qual essa instância deve ser agrupada.

No caso da estimação de densidade por *kernel* gaussiano, a subida de encosta é guiada pelo gradiente da função de densidade, $\hat{f}(X)$, que toma a forma:

$$\nabla \hat{f}(X) = \frac{1}{h^{4n}} \sum_{i=1}^N K(d(X, X_i)/h) \cdot (X - X_i) \quad (3.9)$$

Fazendo o gradiente igual a zero em (3.9) e rearranjando, obtém-se a fórmula iterativa para a posição seguinte na subida de encosta, ou seja, da posição X^k para a posição X^{k+1} , onde X_i representa a posição inicial ($k = 1$):

$$X^{k+1} = \frac{\sum_{i=1}^N K(d(X^k, X_i)/h) X_i}{\sum_{i=1}^N K(d(X^k, X_i)/h)} \quad (3.10)$$

Esse processo iterativo é interrompido quando a variação da função de densidade \hat{f} , de um passo para o outro, seja inferior a um certo limiar ϵ , ou seja:

$$\frac{\hat{f}(X^{k+1}) - \hat{f}(X^k)}{\hat{f}(X^{k+1})} \leq \epsilon \quad (3.11)$$

Assumindo-se um $\epsilon > 0$ apropriado, pode-se esperar que os pontos finais X^N , atingidos por cada instância após N iterações, estejam próximos dos respectivos máximos locais dos grupos aos quais essas instâncias serão agrupadas.

A Figura 3.5 ilustra o método. O campo de densidades é representado por um mapa

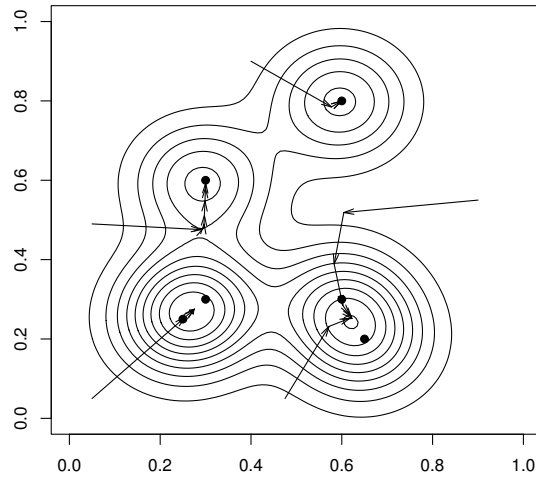


Figura 3.5 - Ilustração do método de subida de encosta, aplicado sobre o campo de densidade de instâncias, que é utilizado no método de agrupamento DENCLUE 2.0. Fonte: [Hinneburg e Gabriel \(2007\)](#)

de contornos, onde podem ser observados os máximos locais. Cada sequência de setas corresponde ao caminho percorrido por uma instância a ser agrupada pela aplicação do método de subida de encosta. Pode-se observar que cada caminho termina no máximo local que representa o grupo ao qual a instância foi agrupada. Pode-se estabelecer um limiar conveniente de densidade para filtrar máximos locais que sejam irrelevantes, isto é, que não estejam associados a um grupo. Finalmente, note-se que o número de grupos depende da largura adotada da janela de suavização.

3.2.2 Árvores de decisão

Métodos de classificação permitem prever a classe de uma instância a partir de seus atributos, após um aprendizado prévio. No caso deste trabalho, um mapa que prevê a ocorrência de chuva forte e convectiva é gerado, sendo adotadas duas classes, no caso, ocorrência ou ausência de chuva forte e convectiva. Cada pixel centralizado num ponto de grade desse mapa corresponde a uma instância, cuja classe indica se ocorrerá ou não chuva forte ou convectiva naquele pixel. Os atributos usados no treinamento e na predição são os índices de instabilidade e as variáveis atmosféricas geradas pelo modelo numérico ETA para cada pixel da grade considerada.

Na fase de treinamento, uma lista de instâncias previamente classificados (conjunto de treinamento) é apresentada ao método, que constrói um modelo, procurando minimizar o erro de classificação. Na fase de validação, procura-se sintonizar parâmetros do método de classificação para buscar um melhor desempenho, usando o conjunto de dados validação. Esses parâmetros podem ser, por exemplo, o nível de

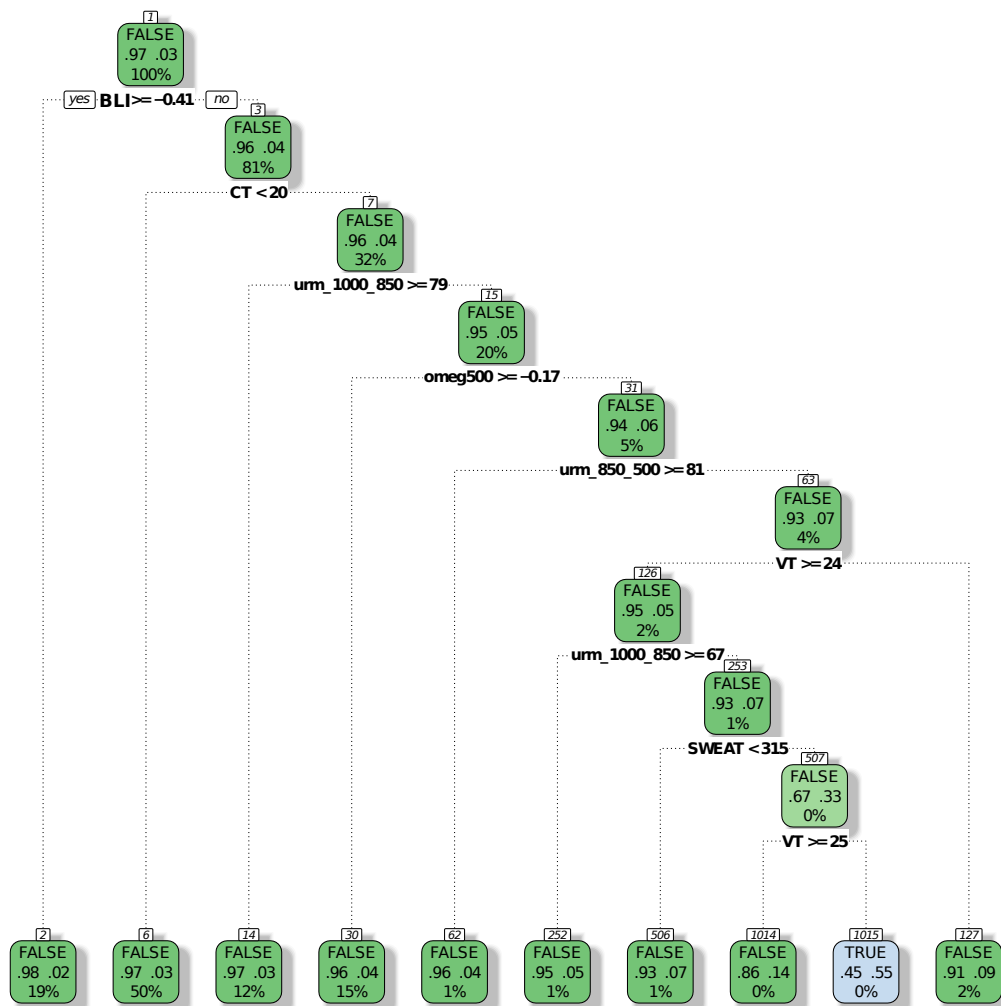


Figura 3.6 - Exemplo de árvore de decisão para verificar ocorrência (TRUE) ou ausência (FALSE) de precipitação forte ou convectiva e forte a partir de índices de instabilidade e variáveis atmosféricas

poda de uma árvore de decisão, ou o número de iterações e o número de neurônios de uma rede neural. Finalmente, o conjunto de teste é usado para estimar o desempenho do classificador. É importante que os conjuntos de teste e validação sejam independentes do conjunto de treinamento, para minimizar o problema do *overfitting*. Esse problema ocorre quando o classificador se torna super-especializado, ou seja, o classificador é capaz de acertar a classe das instâncias que fazem parte do conjunto de treinamento, mas erra ao tentar classificar qualquer instância nova. Caso esse classificador fosse avaliado com base apenas no desempenho do conjunto de treinamento, o resultado seria indevidamente favorável.

Um método de classificação comumente empregado, e adotado neste trabalho, é a árvore de decisão. Optou-se por este método em função de sua simplicidade, de seu uso intuitivo e da facilidade de geração de regras de decisão, as quais podem ser comparadas àquelas da Ferramenta Objetiva de Previsão do CPTEC/INPE. Outro motivo é a facilidade de experimentações decorrentes do ajuste de parâmetros da árvore. No caso, escolheu-se uma árvore de decisão binária do tipo CART (Classification and Regression Tree, (BREIMAN et al., 1984)), que utiliza variáveis contínuas. A árvore de decisão é gerada a partir de um nó raiz. Cada nó interno representa a avaliação de um condicional. Conforme este seja verdadeiro ou falso, é escolhido um dos dois nós descendentes, pois a árvore é binária.

Um exemplo encontra-se na Figura 3.6, no qual ilustra-se a classificação de uma instância correspondente a um pixel na previsão de um modelo numérico como sendo “verdadeira” (com presença de atividade convectiva) ou “falsa” (ausência de atividade convectiva). Este exemplo refere-se a uma árvore que sofreu poda, a título de simplificação. A previsão fornece os atributos de informação usados na classificação. Estes atributos correspondem às variáveis e índices de instabilidade que foram definidos na Seção 2.3 (BLI, CT, etc.).

No mesmo exemplo, o nó raiz contém a avaliação do condicional $BLI \geq -0,41$. Caso essa condição se verifique, segue-se o caminho da esquerda, que não contém nós descendentes (ou seja, não implica em outras avaliações de condicionais), e que leva a um nó terminal com a classificação final atribuindo à instância o valor “falso”, ou seja, não correspondente a chuva convectiva.

Uma árvore de decisão é gerada a partir de um nó raiz que contém todas as instâncias do conjunto de treinamento. A construção da árvore procede selecionando-se, seja para o nó raiz, seja para os nós descendentes, o atributo que melhor divide as instâncias do nó entre as diversas classes, segundo uma métrica de ganho de informação, obtendo-se assim dois nós descendentes com menor impureza. A impureza é uma função que varia no intervalo $[0,1]$, sendo máxima quando o nó apresenta frações iguais de instâncias de cada classe. A geração da árvore objetiva definir nós com baixa impureza, ou seja, que permitam separar as instâncias de cada classe. Em alguns algoritmos de árvore de decisão adota-se a entropia como medida da impureza:

$$E(S) = - \sum_i p_i \log p_i \quad (3.12)$$

onde p_i é a proporção de instâncias da classe i no nó S .

No caso da árvore CART, a medida de impureza é dada pelo índice Gini:

$$Gini(S) = 1 - \sum_i p_i^2 \quad (3.13)$$

O ganho de informação é relativo à escolha de um atributo A para um determinado nó S , sendo dado (para qualquer medida de impureza) pela diferença entre a impureza desse nó e a impureza média dos nós descendentes:

$$Ganho(S, A) = E(S) - I(S, A) \quad (3.14)$$

onde $I(S, A)$ é a impureza média do nó S relativa ao atributo A :

$$I(S, A) = \sum_i \frac{|S_i|}{|S|} E(S_i) \quad (3.15)$$

onde $|S_i|$ é o número de instâncias da classe i do nó S , e $|S|$ é o número total de instâncias do nó S .

Assim, durante o treinamento, a árvore vai sendo construída pela geração de sucessivos nós descendentes até que um critério de parada seja atendido, por exemplo, o número máximo de níveis ou o número mínimo de instâncias por nó. Adicionalmente, pode-se fazer a poda da árvore, percorrendo-se as diversas sub-árvores até se chegar a nós onde o erro de classificação atinja um limiar mínimo, utilizando um conjunto de dados de validação. Isso evita o chamado “overfitting” que pode ocorrer com qualquer algoritmo de classificação o qual se caracteriza por classificar corretamente a maior parte das instâncias do conjunto de treinamento, mas falha em classificar as instância do conjunto de teste.

3.3 Ferramenta Objetiva de Auxílio à Previsão do Tempo

O Grupo de Previsão do Tempo do CPTEC/INPE idealizou uma ferramenta objetiva de auxílio na previsão chamada Ferramenta Objetiva de Previsão do Tempo do CPTEC/INPE que permite visualizar variáveis selecionadas de previsões do modelo numérico regional ETA com resolução de 20 km (ANDRADE et al., 2010), identificando com cores específicas pontos de grade com possibilidade de ocorrência de eventos tais

como pancada de chuva com trovoada, tempestade e granizo. A seleção das variáveis e os limiares adotados para esses três tipos de eventos foi realizada de forma subjetiva com base na experiência dos meteorologistas e em valores de referência citados na literatura da área. As variáveis e limiares formam conjuntos de regras que identificam ou não a ocorrência de um dos tipos de evento extremo para a previsão considerada. Essas visualizações identificam de forma automática a possibilidade de ocorrência desses eventos em dadas áreas, permitindo que o previsor as investigue com mais detalhe e contribuindo para uma melhor previsão de tempo.

As variáveis meteorológicas da Ferramenta Objetiva de Previsão do Tempo do CPTEC/INPE foram selecionadas de forma a ressaltar a combinação das componentes termodinâmica e dinâmica da atmosfera. Algumas variáveis obtidas do modelo foram combinadas na forma de índices de instabilidade. As variáveis adotadas e os respectivos limiares são descritas em (ANDRADE et al., 2010) sendo repetidos abaixo, além da descrição dos três tipos de eventos considerados:

Tipo 1) Pancada de Chuva com Trovoada: evento meteorológico caracterizado por chuvas geralmente moderadas ou intensas e acompanhadas, na maior parte dos casos, de atividade elétrica; a possibilidade de ocorrência deste tipo de evento é sinalizada pela combinação das seguintes variáveis e limiares: umidade relativa média nos níveis de 1000 a 850 hPa e nos níveis de 850 a 500 hPa, acima de 60%; ômega no nível de 500 hPa negativo inferior a -2×10^{-5} Pa/s; índices de instabilidade Total Totals (TT) > 45 e $K > 30$.

Tipo 2) Tempestade: evento meteorológico caracterizado por um maior grau de severidade, geralmente associado com chuvas torrenciais, grandes acumulados de chuva, rajadas de vento, fortes descargas elétricas e até presença de tornados; a possibilidade de ocorrência deste tipo de evento é sinalizada pela combinação das seguintes variáveis e limiares: os mesmos do tipo 1, exceto por $TT > 48$ e $K > 33$.

Tipo 3) Granizo: evento meteorológico similar ao anterior (tempestade), mas com potencial ocorrência de granizo; a possibilidade de ocorrência deste tipo de evento é sinalizada pela combinação das seguintes variáveis e limiares: umidade relativa média na camada 1000/850 hPa e na camada média 850/500 hPa, acima de 60%; ômega negativo em 500 hPa inferior a -3×10^{-5} Pa/s; $SWEAT > 220$, Total Totals (TT) > 52 .

Andrade et al. (2010) apresentaram uma análise qualitativa da Ferramenta Objetiva de Previsão do Tempo do CPTEC/INPE para um evento meteorológico severo

ocorrido na região da tríplice fronteira Brasil-Argentina-Paraguai em setembro de 2009. Pode-se dizer que essa ferramenta apresenta um conjunto de regras para a verificação de ocorrência de cada um dos tipos de eventos extremos considerados. Uma análise quantitativa do grau de acerto das possibilidades mostradas nesses três tipos de visualizações exigiria dados meteorológicos bem detalhados, oriundos de radiossondagem e/ou estações meteorológicas.

3.4 Análise da correlação entre número descargas NS e chuva convectiva

Esta seção apresenta resultados que demonstram a existência de uma correlação entre o número de descargas e chuva produzida por tempestades elétricas, servindo de motivação para este trabalho. Os resultados desta seção são baseadas numa pesquisa paralela, independente desta tese, proposta por [Garcia et al. \(2013\)](#).

Dados de radar meteorológico e dados de descargas de 2009 foram compilados para uma área quadrada A de 50 km de lado no noroeste da cidade de Bauru (a cidade fica exatamente no vértice sudeste do quadrado). Dados de radares meteorológicos fornecidos por radares de Bauru e Presidente Prudente (descritos na Seção 2.2) foram utilizados para estimar a precipitação convectiva, por meio de uma relação $Z-R$ adequada ([CALHEIROS; GOMES, 2010](#)). Foram produzidas séries temporais de descargas NS acumuladas em 30 min, assim como de precipitação convectiva, durante três períodos de tempo, sempre considerando essa área quadrada: o ano de 2009, o mês de setembro de 2009, e 54 horas de 08 de setembro a 10 de setembro de 2009, para cobrir a extensão de uma tempestade particular. As curvas que correspondem à série temporal desses períodos de tempo aparecem nas Figuras 3.7, 3.8 e 3.9, as quais foram obtidas na pesquisa relacionada de [Garcia et al. \(2013\)](#). Estas curvas foram suavizadas por um filtro Gaussiano unidimensional. A largura do filtro é de 24 horas (48 amostras) para o primeiro conjunto de curvas (ano de 2009), 2 horas (quatro amostras) para o segundo (setembro de 2009), e uma hora para a terceira (tempestade de 8 a 10 de setembro de 2009).

Os valores de correlação entre o número de descargas NS e a massa precipitada foram então calculados para os três conjuntos de curvas, mas usando os dados brutos, antes da suavização. No entanto, o cálculo de correlação usando séries temporais completas seria inútil, já que a maioria dos pontos de dados não apresentam chuva ou descargas. Portanto, foram identificadas 259 tempestades elétricas para o ano de 2009 e 31 tempestades elétricas para o mês de setembro de 2009, sempre para a mesma área quadrada. Um evento foi considerado como uma tempestade se apresentou chuvas acima de um limite, se as chuvas persistiram por pelo menos dois

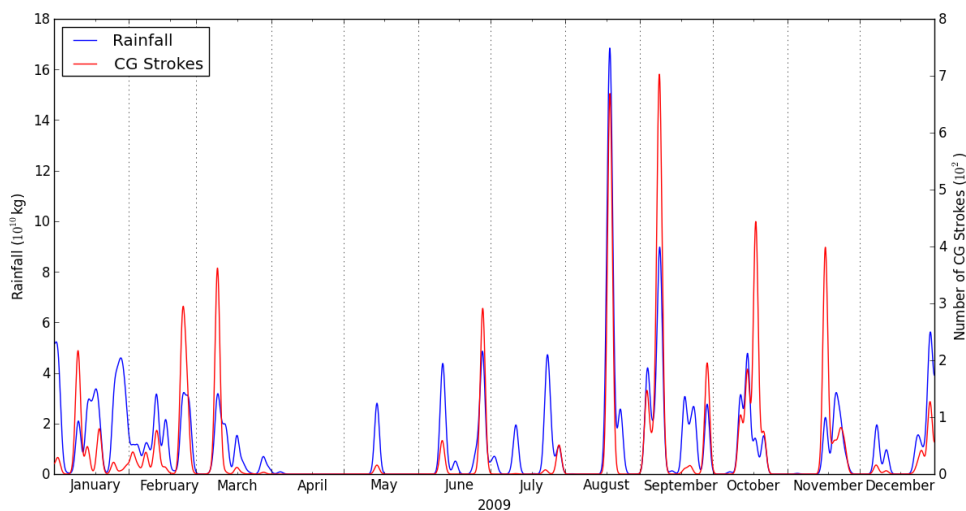


Figura 3.7 - Série temporal dos valores acumulados de descargas NS em 30 min e a massa precipitada (chuva) para ano de 2009 dentro da área A de 2.500 km² próxima Bauru (curvas suavizadas por um filtro gaussiano unidimensional).

intervalos de 30 min e se também cumpriu com os critérios de Steiner para a precipitação convectiva. A correlação é então calculada entre o número de descargas NS e a massa precipitada para cada tempestade, sempre considerando os valores acumulados de 30 min do número de descargas e de chuva. Também foram realizados alguns testes de correlação cruzada, utilizando os mesmos dados, mas a correlação máxima foi obtida com atraso zero.

O primeiro caso refere-se às 259 tempestades elétricas incluídas na série temporal do número de descargas NS acumuladas em 30 min e da massa precipitada para todo o ano de 2009 (ver Figura 3.7). Valores de correlação foram calculados para cada uma das 259 tempestades e apresentou média de 0,77, mediana de 0,78, com desvio padrão de 0,15, valor mínimo de 0,29 e máximo de 0,98.

O segundo caso refere-se às 31 tempestades incluídas na série temporal do número acumulado de descargas NS em 30 min e da massa precipitada para o mês de setembro de 2009 (ver Figura 3.8). Valores de correlação foram calculados para cada uma das 31 tempestades. Estes valores ficaram no intervalo [0,53;1,00], apresentando média e mediana de 0,82, com desvio padrão de 0,12.

Finalmente, o terceiro caso se refere a uma única tempestade que apareceu na série temporal do número acumulado de descargas NS em 30 min e da massa precipitada entre os dias 8 a 10 de setembro de 2009 (veja a Figura 3.9). A correlação calculada

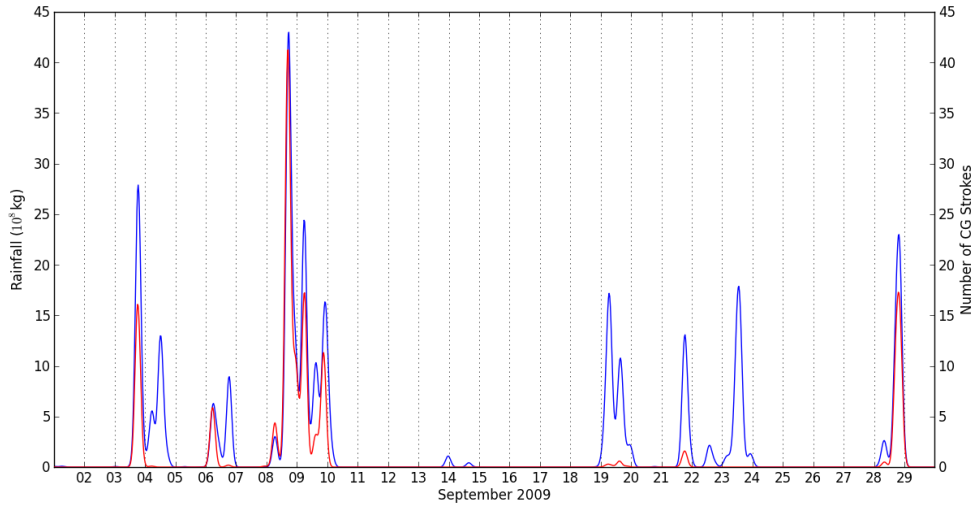


Figura 3.8 - Série temporal dos valores acumulados de descargas NS em 30 min e a massa precipitada (chuva) para o mês de setembro de 2009 dentro da área A de 2.500 km^2 próxima a Bauru (curvas suavizadas por um filtro gaussiano unidimensional).

para esta tempestade em particular é 0,78.

3.5 Estimação da precipitação acumulada a partir de descargas elétricas atmosféricas

Garcia et al. (2013) estimaram uma função que correlaciona a precipitação acumulada e o número de descargas NS num determinado período de tempo e para uma área ao redor dos radares meteorológicos de Bauru e Presidente Prudente. Tipicamente, essa correlação é expressa pela razão entre precipitação e descargas (RLR), expressa por um valor constante (TAPIA et al., 1998). Esta função foi estimada por meio de uma janela deslizante temporal, sendo expressa por:

$$WRLR(N) = a \times b^N + c \quad (3.16)$$

onde WRLR (*windowed RLR*) é a estimativa de precipitação convectiva em 10^6 kg, N é o número de descargas, e $\{a, b, c\}$ são parâmetros do modelo, sendo seus valores, respectivamente, 941,3, 0,3878, e -182,1 (a e c são dados em 10^6 kg).

Constatou-se que o uso de setores quadrados de 50 km de lado e de intervalos de tempo de 30 min maximizavam a correlação entre precipitação e descargas NS.

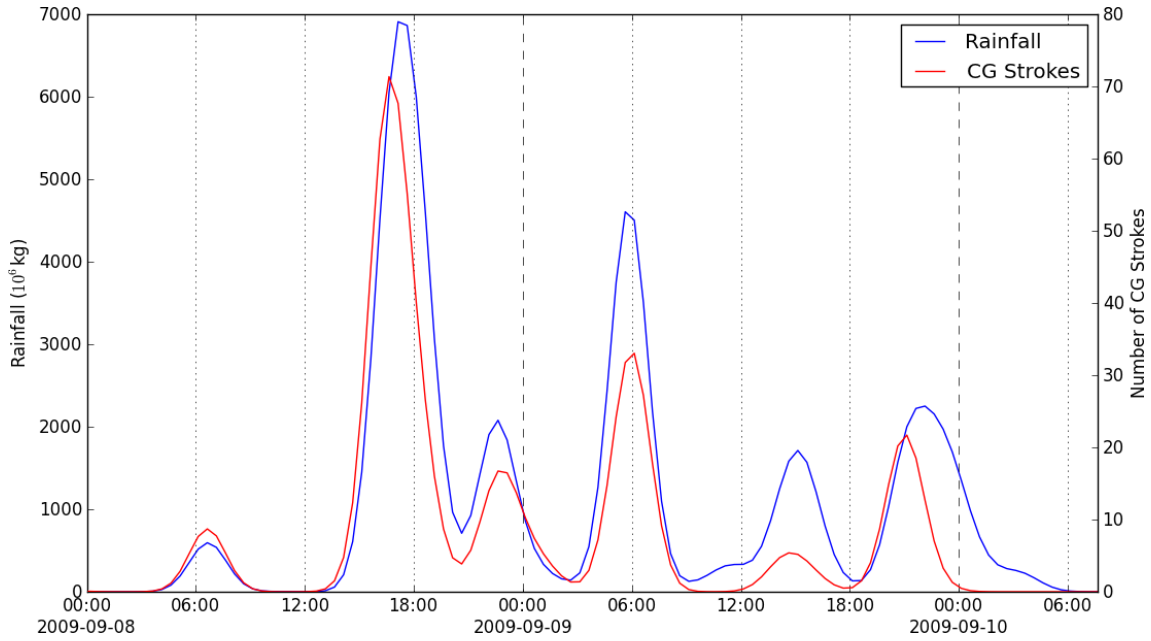


Figura 3.9 - Série temporal dos valores acumulados de descargas NS em 30 min e a massa precipitada (chuva) para alguns dias no mês de setembro de 2009 dentro da área A de 2.500 km² próxima a Bauru durante uma tempestade em particular (curvas suavizadas por um filtro gaussiano unidimensional).

Assumiu-se também que, dentro de cada quadrado de 50 km de lado, a precipitação distribui-se de maneira proporcional à densidade de descargas normalizada. Isso permite a geração de mapas de precipitação acumulada com resoluções da ordem de quilômetros

A abordagem proposta em Garcia et al. (2013) introduziu as seguintes inovações: a massa precipitada é estimada por uma função não-linear, em vez de uma constante multiplicada pelo número de descargas; o uso de uma janela deslizante no tempo na estimação da função, e obtenção da distribuição espacial da precipitação a partir do campo de densidade de descargas NS.

Numa evolução posterior desse trabalho, a função original foi extrapolada para quadrados de 0,5° de lado fora da cobertura do radar. Isso implicou no cálculo de coeficientes de ajuste K_{ij} para os quadrados que abrangem a região de cobertura da RINDAT. Estes coeficientes foram calculados com base em médias trimestrais de precipitação e de descargas NS para cada quadrado ij .

3.6 Critério de identificação de chuva convectiva em imagens de radar

Um critério para identificar a chuva convectiva em imagens de radar meteorológico foi proposto em Steiner et al. (1995). O critério identifica um pixel como apresentando chuva convectiva numa imagem CAPPI 3 km nos seguintes casos:

- a) Qualquer pixel acima de 40 dBZ é considerado um centro convectivo;
- b) Qualquer pixel que exceda a refletividade média Z_m (em dBZ) de um valor ΔZ , considerando-se num círculo de 11 km de raio centrado no mesmo, também é considerado um centro convectivo, onde $\Delta Z(\text{dB}) = 10 - Z_m^2/180$;
- c) Os pixels que compõem a vizinhança circular em torno de um centro convectivo de raio R também são considerados convectivos, conforme a equação 3.17 abaixo (os colchetes incompletos denotam que se considera apenas a parte inteira da fração):

$$R(\text{km}) = \lfloor \frac{Z_m - 15}{5} \rfloor - 1, \quad 1\text{km} \leq R \leq 5\text{km}. \quad (3.17)$$

4 OS SOFTWARES EDDA E EPPA

Neste capítulo são apresentados os softwares EDDA e EPPA juntamente com os métodos específicos e implementações associados. Alguns outros métodos utilizados já foram descritos no Capítulo 3.

A Seção 4.1 descreve o software EDDA que gera campos de densidade de descargas elétricas atmosféricas que já vem sendo avaliado operacionalmente no CEMADEN e que utiliza o método de estimação de densidade descrito na Seção 3.1. A Seção 4.2 descreve a implementação futura do software EDDA com o agrupamento espaço-temporal de descargas. A Seção 4.3 introduz o software EDDA com estimação de precipitação acumulada de origem convectiva, utilizando o método descrito na Seção 3.5. Finalmente, a Seção 4.4 aborda o software EPPA de previsão de ocorrência de precipitação a partir de dados de modelo de previsão meteorológica.

4.1 O software EDDA

O software EDDA estima a densidade de ocorrência de descargas elétricas atmosféricas para uma extensão geográfica e intervalo de tempo selecionados (STRAUSS et al., 2010). O campo de densidade de ocorrências é suave, delimitando mais claramente a região de atividade convectiva a partir das descargas, as quais são muito esparsas no espaço e no tempo. A aplicação dessa abordagem foi proposta originalmente em Politi et al. (2006) com o intuito de rastrear a atividade convectiva eletricamente ativa por meio das descargas NS.

Os dados brutos de descargas, contendo os registros individuais em formato ASCII são gerados pela Rede Integrada Nacional de Detecção de Descargas Atmosféricas (RINDAT), fornecidos pelo CPTEC/INPE. São gerados arquivos em formato ASCII, adequados a algoritmos de mineração, e em formato de grade binário, por exemplo, para a ferramenta de visualização meteorológica GRADS.

O software EDDA implementa o estimador de *kernel* gaussiano com uma largura de janela fixa (Equação 3.5) ou automática (Equação 3.8). A Terra é considerada como uma esfera e a distância $d(A, B)$ sobre a superfície entre dois pontos A e B é aproximada pelo cálculo de uma distância euclidiana, porém corrigindo-se a componente ao longo das latitudes pelo cosseno da latitude média dos pontos considerados na imagem, uma vez que esta distância não está sobre um círculo máximo da superfície terrestre. Na equação abaixo, “lat” e “lon” denotam as latitudes e longitudes, respectivamente, enquanto que “lat_{max}” e “lat_{min}”, a latitude máxima e mínima dos

pontos de grade considerados.

$$d(A, B)^2 = (\text{lat}_A - \text{lat}_B)^2 + (\text{lon}_A - \text{lon}_B)^2 \times \cos\left(\frac{\text{lat}_{max} + \text{lat}_{min}}{2}\right) \quad (4.1)$$

O software EDDA basicamente calcula a densidade de descargas em cada ponto da grade definida pelo usuário. Numa primeira implementação do software, a rotina específica para esse cálculo possuía um laço (*loop*) que varria todos os pontos de grade e, para cada um destes, outro laço varria todas as descargas da lista gerada a partir dos dados de entrada. Considerando-se uma grade quadrada com N pontos de lado e uma lista com N_R descargas, a complexidade algorítmica era $O(N^2 \times N_R)$. O desempenho computacional da rotina era otimizado pela não computação da influência de descargas acima de certa distância no cálculo da densidade resultante em cada ponto de grade. A versão atual foi inspirada em algoritmos para resolução de problemas N-corpos, especificamente na divisão do domínio espacial em caixas (*bins*). Por conveniência (vide Seção 4.3 adiante), a grade foi dividida em caixas de $0,5^\circ \times 0,5^\circ$ sendo computados para o cálculo da densidade em cada ponto de grade da caixa apenas as descargas que ocorreram dentro da caixa, para o intervalo de tempo considerado, e aquelas que ocorreram nas caixas vizinhas, compondo assim uma vizinhança de 3×3 (9 caixas). O desempenho computacional foi assim otimizado por um fator da ordem de 200 (para a grade e resolução utilizadas pelo CEMADEN). Esse fator resulta aproximadamente da razão entre as 9 caixas de cada vizinhança e o total de caixas que abrangem toda a grade. A complexidade algorítmica da versão atual é, portanto, semelhante à da versão original. Entretanto, a versão atual implementa de maneira mais eficiente a limitação das descargas distantes do ponto de grade considerado, enquanto que na versão original, necessita-se calcular a distância de cada descarga ao ponto de grade e verificar se esta excede o limiar adotado. Considerando-se a alta resolução espacial e a grande extensão da grade adotada pelo CEMADEN (Seção 5.1), a versão atual tem desempenho computacional suficiente, prescindindo de uma eventual paralelização. A versão original chegou a ser facilmente paralelizada com o uso de *threads*, mas mesmo assim fica aquém da versão atual, que é sequencial.

O software EDDA emprega um arquivo de configuração para selecionar o intervalo de tempo, a área geográfica, o tamanho da grade, bem como um filtro para selecionar a polaridade e o tipo das descargas. É possível selecionar saídas no formato de tabelas ASCII contendo a longitude, latitude e a densidade calculada em ponto de grade, objetivando seu uso na mineração de dados meteorológicos existentes. O

software permite também visualizar animações varrendo os registros de descargas por meio de uma janela deslizante, de forma a acompanhar a evolução temporal de estruturas convectivas. A pronta disponibilidade de dados de descargas torna possível gerar animações que permitem monitorar os eventos meteorológicos com atividade elétrica em tempo quasi-real, possibilitando ao meteorologista uma visão instantânea de estruturas convectivas e de sua evolução recente. Assim, o software proposto tem potencial operacional em Meteorologia.

Adicionalmente, o software EDDA já foi utilizado em alguns trabalhos relativos à mineração de dados meteorológicos, no caso, utilizando valores altos de densidade de descargas como atributo de decisão indicativo de atividade convectiva. O campo de densidade de ocorrência de descargas elétricas atmosféricas NS foi usado na fase de treinamento de algoritmos que buscam identificar padrões em previsões de modelos numéricos (LIMA; STEPHANY, 2013a; LIMA; STEPHANY, 2013b; PESSOA et al., 2012).

4.2 O software EDDA com agrupamento espaço-temporal de descargas

Uma nova funcionalidade foi desenvolvida para o software EDDA, incorporando o agrupamento espaço-temporal de descargas NS para identificar células eletricamente ativas, seguido da estimação da densidade de descargas (STRAUSS; STEPHANY, 2011; STRAUSS et al., 2013). Dessa forma, células eletricamente ativas (correspondentes a cada grupo) podem ser identificadas e monitoradas ao longo do tempo. Adicionalmente, torna-se possível obter parâmetros de cada célula eletricamente ativa, como o número total de descargas num intervalo de tempo ou a taxa de ocorrência de descargas. Essa funcionalidade ainda não foi implementada no EDDA mas poderá ser alcançada com relativa facilidade.

Um fluxograma da implementação atual do EDDA, que permite a exibição de imagens estáticas e animações, é mostrado na Figura 4.1 (à esquerda), enquanto que o fluxograma da implementação operacional futura com agrupamento é mostrado na Figura 4.1 (à direita). Os dois fluxogramas são explicados a seguir.

Na versão atual do EDDA, o software fica em modo de espera, monitorando a chegada de novos arquivos que contém dados de descargas. O intervalo de tempo selecionado pelo usuário e a área permite ao software gerar o campo de densidade de ocorrência de descargas NS para a área e instante de tempo considerados, por meio da estimação de densidade. A imagem correspondente, obtida a partir de um mapa de contornos e uma escala de cores, é a imagem atual. O software monitora automaticamente a chegada de novos dados para compor os quadros subsequentes. O

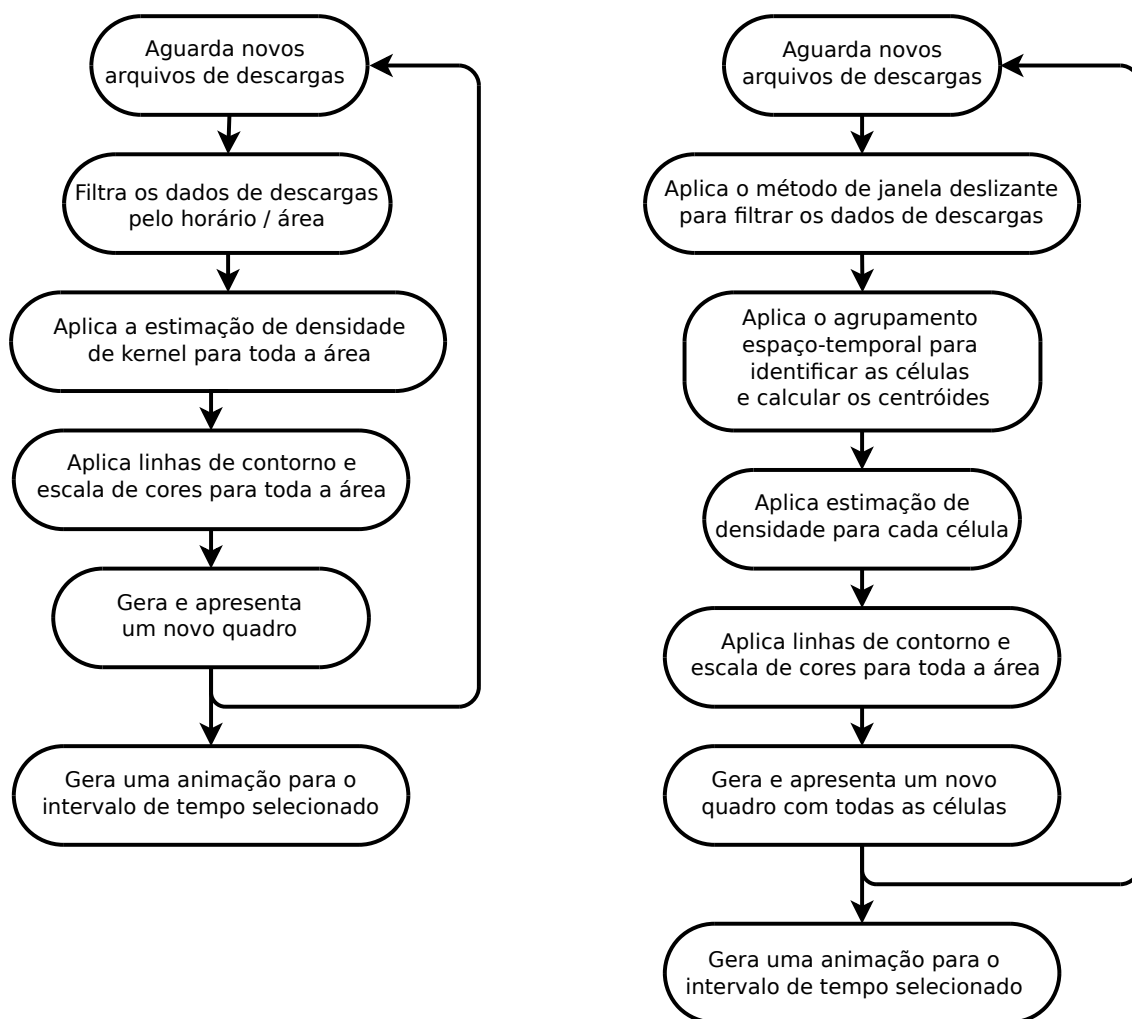


Figura 4.1 - Fluxograma das implementações operacionais do software de EDDA (à esquerda) e o mesmo software com agrupamento de descargas (à direita).

rastreamento de uma célula específica é realizado visualmente, selecionando quadros sucessivos. O usuário também pode definir um intervalo mais longo de tempo (por exemplo, algumas horas) para ver uma animação mostrando a evolução da atividade de raios nas últimas horas. O sistema armazena quadros gerados anteriormente para eventualmente gerar animações.

O software EDDA com agrupamento espaço-temporal de descargas baseia-se em uma janela deslizante temporal, semelhante à empregada para controle de fluxo de dados em redes (STENNING, 1976), mas empregada de maneira inédita aqui. Uma janela temporal de largura fixa no tempo se desloca com uma velocidade constante, mas em intervalos de tempo discretos, esperando pela chegada de novos dados. Os agrupamentos resultantes correspondem às células eletricamente ativas. Uma

vantagem adicional do agrupamento espaço-temporal é que uma lista de todas as descargas que compõem cada agrupamento se torna disponível.

Uma grande dificuldade neste processo de agrupamento é identificar e rastrear uma célula específica ao longo do espaço e tempo mantendo sua identidade, e a manipulação de divisões e combinações de células. Depois de realizar o agrupamento 2D dos dados de raios em um determinado intervalo de tempo, os grupos resultantes são uma mistura de novos grupos e aqueles identificados no intervalo de tempo anterior. Aqui, o avanço da janela foi selecionado como metade do intervalo de tempo. Consequentemente, todas as descargas no intervalo de tempo considerado são atualizadas após dois avanços da janela deslizante. O algoritmo mantém uma lista de descargas para cada célula. As células têm identificadores. Um grupo pode persistir se novas descargas são incluídas em sua lista de descargas. Se uma célula divide-se eventualmente, o fragmento com mais descargas preserva sua identidade original, enquanto um novo identificador é atribuído para cada um dos segmentos restantes. Também podem se juntar duas células, assumindo a identidade daquela com mais descargas. Um grupo também pode desaparecer, se não houver novas descargas para ser anexadas a ele.

Por exemplo, como pode ser visto na Figura 4.2, dois grupos aparecem no intervalo de tempo T , rotulados como #1 e #2. Na próxima iteração $T + \Delta T$, quatro estruturas aparecem, mas correspondendo aos grupos #1, #2, #3 e #4. Duas novas descargas formaram o novo grupo #3, já que elas não estavam perto dos agrupamentos existentes. A estrutura do quarto agrupamento é devida à separação do agrupamento #2: o fragmento esquerdo tem preservada sua identidade, já que tem mais descargas, mas o fragmento a direita forma o novo grupo #4. A contagem de descargas do grupo #1 aumentou, já que novas descargas foram anexadas ao arquivo. A janela deslizante permite descartar dados antigos, enquanto adiciona os novos dados a um ritmo constante. Portanto, os grupos podem desaparecer, juntarem-se ou dividirem-se ou, ainda, grupos novos podem ser criados.

Os parâmetros desse método que precisam ser ajustados são a largura da janela e o intervalo temporal do deslizamento. A largura da janela deve ser grande o suficiente para fornecer uma amostragem adequada, mas não muito grande para evitar a perda de resolução temporal. O intervalo temporal para o deslizamento deve ser suficientemente pequeno para permitir uma correlação significativa entre os grupos, mas não demasiado pequeno para evitar excessivo tempo de processamento.

O algoritmo de agrupamento escolhido é DENCLUE 2.0, baseado em estimação de

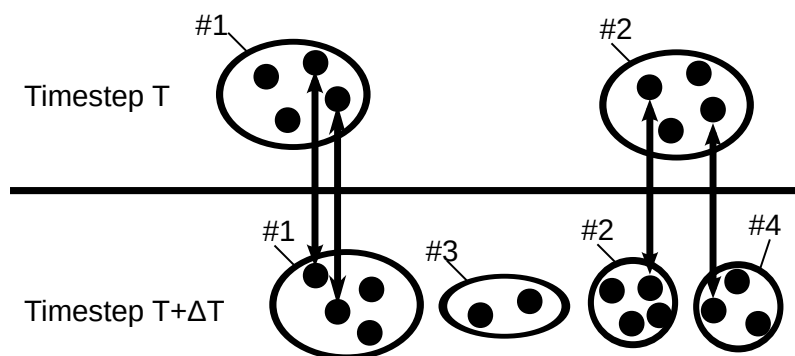


Figura 4.2 - Um exemplo da evolução do agrupamentos de descargas de um intervalo de tempo para outro, de acordo com o critério de fusão/separação adotado.

densidade (HINNEBURG; GABRIEL, 2007). Este algoritmo foi descrito na Seção 3.2.1. O uso de todas as descargas como sementes por DENCLUE dá-lhe robustez com relação à escolha das sementes iniciais. Neste sentido, é semelhante ao método de Goodman (1990) de atribuir os máximos locais do campo de densidade como as sementes iniciais do método de *k-means*. Por outro lado, os algoritmos baseados em *k-means* não podem separar agrupamentos próximos com fronteiras intrincadas (não-lineares) (HAN; KAMBLER, 2011), enquanto isso não é uma limitação do DENCLUE.

No caso da abordagem proposta, o software EDDA (Figura 4.1) também espera a chegada de novos arquivos de dados de raios e seleciona o intervalo de tempo (10 min) e a área. No entanto, as descargas são filtradas pela janela deslizante e submetidas ao agrupamento espaço-temporal que produz uma lista de grupos que correspondem às células eletricamente ativas. A estimação de densidade é então aplicada para cada grupo separadamente, resultando em um conjunto de campos de densidade de ocorrência de descargas NS. Um gráfico de contornos e uma escala de cores são empregado para todas as células da imagem (*frame*) obtida. O rastreamento de uma célula específica pode ser realizado visualmente, selecionando quadros sucessivos, ou mediante as sucessivas posições do centroide da célula. Animações também podem ser produzidas.

A rede RINDAT fornece dados de descargas elétricas atmosféricas com resolução de milissegundos, mas tais dados são atualizados a cada 5 min e, portanto, a janela deslizante é também avançada sempre de 5 min. O esquema de sobreposição da janela deslizante requer que o valor da largura da janela seja maior do que 5 min, para não ter espaços vazios nos dados. Portanto, foi escolhida uma largura de janela de 10 min, resultando em uma sobreposição de 50% entre uma janela e a próxima. Uma janela maior não seria conveniente para fins de previsões, e uma menor iria exigir

muito processamento, além de não ser múltiplo do intervalo de 5 min. Outra razão foi ter compatibilidade com a resolução dos dados de radares meteorológico disponíveis, que foram interpolados em 10 min, uma vez que suas resoluções temporais variam entre 7,5 e 10 min, de acordo com o radar.

4.3 O software EDDA com estimação de precipitação

O software EDDA incorporou uma nova funcionalidade, a estimação da precipitação acumulada a partir de dados de descargas NS, conforme descrito na Seção 3.5. A nova versão EDDA-chuva, que incorpora ao EDDA a estimação de precipitação a partir de dados de descargas NS incorpora as matrizes que contém os coeficientes K_{ij} para os quatro trimestres do ano. A cada 30 min, o EDDA-chuva acumula as descargas que ocorreram em cada quadrado ij de $0,5^\circ \times 0,5^\circ$ e obtém a massa precipitada nesse intervalo de tempo pela aplicação da Equação 3.16 corrigida pelo coeficiente K_{ij} em cada quadrado para obter a quantidade de chuva acumulada nesse setor. A distribuição espacial da precipitação estimada no quadrado é obtida pela correspondente densidade de descargas NS, sendo expressa em mm de precipitação. O valor obtido é multiplicado pela densidade normalizada de descargas dentro desse quadrado, e o resultado convertido em mm. A cada hora, é gerada uma saída com a soma das saídas de 30 min, sendo assim possível obter acumulados de precipitação para períodos maiores, bem como visualizar a correspondente distribuição espacial da precipitação.

O EDDA-chuva foi desenvolvido com base no software EDDA, acrescentando a funcionalidade de estimação de precipitação, tendo já sido testado e avaliado. Sua avaliação operacional deverá se iniciar em breve no CEMADEN e também no DSA/CPTEC. Seu desempenho computacional é semelhante ao EDDA, conforme exposto na Seção 4.1, uma vez que as estimativas de precipitação são geradas somente a cada 30 min e para uma grade com resolução espacial menor ($0,1^\circ$).

4.4 O software EPPA

O software EPPA é constituído por um classificador que tem por objetivo prever chuva convectiva e forte utilizando dados de previsão numérica e análise. Este software foi inspirado na Ferramenta Objetiva de Previsão do Tempo do CPTEC/INPE (ANDRADE et al., 2010), descrita anteriormente na Seção 3.3. O software EPPA utiliza os mesmos índices de instabilidade, derivados do modelo ETA, mas utiliza um método de aprendizagem de máquina, no caso uma árvore de decisão, para gerar novas regras a partir da precipitação estimada pelo radar. No caso, a árvore de de-

ção adotada foi a árvore CART, abordada na seção 3.2.2, e que integra o pacote RPART (THERNEAU et al., 2013), o qual faz parte do ambiente de software gratuito R (R CORE TEAM, 2013) para computação estatística e gráfica. Embora esta árvore de decisão seja específica para variáveis contínuas, sua implementação pressupõe a adoção de uma discretização específica para cada variável utilizada. Esta discretização não é transparente ao usuário. Este ambiente de software possibilita diversas experimentações numéricas com o intuito de otimizar o desempenho de classificação da árvore, conforme será descrito adiante nos resultados correspondentes. Futuramente, para fins de implementação do software EPPA, pretende-se portar a árvore otimizada ou conjunto de árvores otimizadas para a linguagem C.

Strauss et al. (2012) aplicaram uma árvore CART do pacote RPART para prever a ocorrência de chuva, convectiva ou não, com qualquer intensidade.

A abordagem foi aprimorada para previsão de chuva forte e convectiva. Inicialmente, identificaram-se os pixels das imagens de radar onde ocorreu chuva forte ou convectiva. Precipitação acima de 7,6 mm foi considerada como forte (AMERICAN METEOROLOGICAL SOCIETY, 2013), enquanto que a chuva convectiva foi estimada pelo critério de Steiner (STEINER et al., 1995).

O passo seguinte foi reamostrar os dados de radar na mesma grade espacial dos dados do modelo. Como há uma diferença entre as resoluções do modelo numérico ETA 5 km (0,05°) e dados de radar (0,02°), existem cerca de 3×3 ou 4×4 pixels da grade do radar parcialmente ou totalmente cobertos por um único pixel do modelo. Esse esquema é apresentado na Figura 4.3. Os pixels maiores, em vermelho, correspondem, aos pixels do modelo ETA 5 km. Os pixels menores, em preto, são os pixels do radar. O pixel #1 cobre 9 pixels de radar (3×3), os pixels #2 e #3 cobrem 12 pixels (4×3 e 3×4), enquanto o pixel 4 cobre 16 pixels (4×4).

O critério adotado para considerar um determinado pixel do modelo como apresentando chuva forte ou convectiva é que pelo menos um dos pixels do radar contidos nesse pixel apresentem chuva forte ou convectiva.

Para comparar os valores previstos (presença de chuva convectiva ou chuva forte) e observados, foi adotada a seguinte métrica (NURMI, 2003):

$$\text{Probabilidade de detecção (POD)} = \text{VP} / (\text{VP} + \text{FN}) \times 100\% \quad (4.2)$$

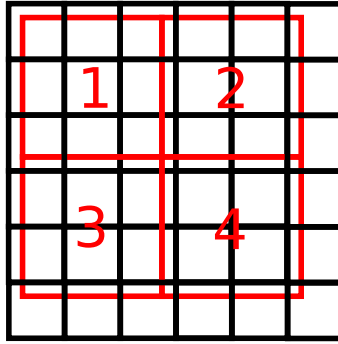


Figura 4.3 - Comparação entre a grade de radar ($0,02^\circ$) e a grade do modelo ETA ($0,05^\circ$).

$$\text{Razão de falso alarme (FAR)} = \text{FP} / (\text{VP} + \text{FP}) \times 100\% \quad (4.3)$$

onde os “positivos” ($\text{VP} + \text{FN}$) são os pixels onde forte ou chuva convectiva foi estimada, “previstos” ($\text{VP} + \text{FP}$) referem-se aos pixels onde as regras da árvore foram satisfeitas, verdadeiros positivos (VP) são os pixels previstos corretamente como positivos, falsos positivos (FP) são os pixels previstos incorretamente como positivos, falsos negativos (FN) são os pixels previstos incorretamente como negativos e verdadeiros negativos (VN) são os pixels previstos corretamente como negativos.

Deve-se ressaltar que, no contexto de previsão, os falsos negativos são mais indesejáveis que os falsos positivos, uma vez que a não previsão de uma ocorrência de chuva forte ou convectiva é mais crítica que um falso alarme.

No entanto, a utilização de modelos numéricos pode subestimar os verdadeiros positivos e superestimar os falsos positivos, devido a sua capacidade limitada em prever, com exatidão, o posicionamento dos sistemas de precipitação de tamanho da mesoescala ou menores. Propõe-se o conceito de “vizinhanças relaxadas”, que consideram um pixel positivo como um verdadeiro positivo se houver uma previsão positiva perto do pixel, mesmo se o pixel em si não foi previsto como positivo. Da mesma forma, se houver um pixel positivo perto de um pixel predito, não é considerado como um falso positivo, mesmo se o pixel em si não é positivo.

A vizinhança adotada são todos os pixels adjacentes ao pixel considerado (horizontalmente, verticalmente ou diagonalmente), resultando numa vizinhança espacial de 15x15km. Também, uma vizinhança temporal é adotada: os positivos podem ocorrer até 30 min antes ou após a previsão.

5 RESULTADOS

Os softwares EDDA e EPPA foram testados para alguns eventos selecionados de atividade convectiva. A validação desses softwares somente ocorrerá com a avaliação em ambiente operacional pelo CEMADEN e, possivelmente, pelo DSA/CPTEC/INPE. Como resultado dessa avaliação, eventuais parâmetros desses softwares, ou mesmo metodologias integrantes, poderão ser revistos. Esse processo iterativo de experimentação e teste é típico em mineração de dados, no qual diversas opções de pré-processamento de dados e de ajuste de parâmetros dos algoritmos são progressivamente refinados. Assim, otimiza-se o processo de mineração de dados visando obter visualizações e/ou classificações que permitam identificar ou prever a ocorrência de atividade convectiva.

É importante notar que os resultados aqui expostos referem-se a estudos de caso selecionados para validar esses softwares, os quais foram publicados, total ou parcialmente, nas referências citadas. Não existe portanto uma compilação mais extensa de resultados de testes referentes a esses softwares. Entretanto, espera-se obter uma compilação mais extensa de resultados com a avaliação operacional desses softwares no CEMADEN.

Este trabalho contou com dados de diversas naturezas, que foram obtidos graças ao CEMADEN, ao CPTEC/INPE, ou ao IPMET/UNESP. Entretanto esses dados sofreram diversas limitações. Dados de radares meteorológicos somente estão disponíveis dentro da sua restrita cobertura, comparativamente à grande extensão do território brasileiro. Dados de descargas elétricas atmosféricas tem tido sua abrangência ampliada nos últimos anos, graças à expansão da rede de sensores de descargas. Por exemplo, dados da rede BrasilDAT deverão estar disponíveis em breve, ampliando consideravelmente a cobertura em relação à rede RINDAT, que gerou os dados utilizados neste trabalho. A rede BrasilDAT deverá disponibilizar também dados de descargas IN. No caso de dados de modelos, utilizaram-se inicialmente dados do modelo ETA 20 km (resolução de $0,2^\circ$) para fazer a predição de ocorrência de qualquer chuva (convectiva ou não e de qualquer intensidade). Entretanto, essa resolução mostrou-se insuficiente para prever especificamente chuva convectiva e forte, demandando dados do ETA 5 km, que tem resolução de $0,05^\circ$. Este último modelo continua sendo executado pelo CPTEC/INPE, mas com abrangência restrita à Serra do Mar paulista, e não mais abrangendo a área mais extensa dos dados deste trabalho, que compreendem todo o estado de São Paulo. Tais restrições afetam o desempenho de abordagens de mineração de dados, que depende da disponibilidade

de bases de dados extensas em termos espaciais e temporais.

Inicialmente, a Seção 5.1 aborda o software EDDA sem agrupamento de descargas: a visualização do campo de densidade de ocorrência de descargas elétricas atmosféricas NS na qual assume-se que as regiões mais densas correspondam a atividade convectiva. Essa visualização é comparada com visualizações de atividade convectiva obtidas a partir de outros dados (radar, satélite, etc.) no CEMADEN. Em seguida, a Seção 5.2 ilustra a aplicação do software EDDA com agrupamento de descargas na geração de campos de densidade de descargas elétricas atmosféricas para algumas tempestades selecionadas, além da avaliação quantitativa da correlação entre descargas e chuva convectiva para esses eventos. Os resultados com e sem o uso de agrupamento espaço-temporal de descargas são comparados. Também, as células individuais dessas tempestades são identificadas e rastreadas. Para algumas dessas células, a evolução temporal da atividade elétrica e da chuva é apresentada.

A seguir, os resultados do software EPPA são apresentados. Na Seção 5.3, após testes preliminares usando a Ferramenta Objetiva de Previsão do Tempo do CP-TEC/INPE, os dados do modelo ETA 20 km são usados para treinar uma árvore de decisão que procura prever a ocorrência de chuva em qualquer quantidade. Os resultados com e sem o uso de vizinhança são comparados. O desempenho global para o mês de janeiro de 2010 é avaliado, e mapas de ocorrência de chuva para algumas tempestades em particular são apresentados como ilustração. A necessidade de melhorar os resultados para chuva convectiva levou a novos testes utilizando dados de maior resolução espacial e temporal. Foi adotado assim o modelo ETA 5 km. Os resultados do uso desse modelo para previsão de chuva forte e convectiva são também apresentados nessa sessão.

5.1 Uso do software EDDA para monitoramento de atividade convectiva

O software EDDA gera campos de densidade de ocorrência de descargas elétricas atmosféricas. Este software vem sendo avaliado operacionalmente no CEMADEN desde outubro de 2012, na versão sem agrupamento espaço-temporal de descargas. O CEMADEN desenvolveu e utiliza o ambiente SALVAR - Plataforma de Monitoramento e Alertas de Desastres Naturais. Essa plataforma permite a visualização e manipulação de dados geoespaciais para subsidiar a tomada de decisões do grupo operacional relativamente à emissão de alertas para a Defesa Civil. Esses alertas são disparados em função de volumes grandes de precipitação acumulada ou então do aumento da vazão de rios em bacias hidrográficas, cenários com potencial de causar inundações ou deslizamentos de terra. Esse ambiente integra dados diversos,

incluindo contornos extraídos de mapas, imagens de radares meteorológicos e de satélites, além de outras informações tais como as referentes ao tipo de ocupação da terra. Permite também integrar dados pontuais, como no caso de dados de estações meteorológicas.

O grupo operacional do CEMADEN gerou algumas imagens, nas quais se compara a visualização de eventos convectivos obtidas pelo software EDDA (áreas em azul claro, azul escuro e verde), sobreposta às demais visualizações usualmente adotadas. A Figura 5.1 mostra a imagem de satélite GOES na banda infravermelha, canal 4, pseudocolorida para mostrar a temperatura de brilho do topo das nuvens. A parte realçada (verde) corresponde às temperaturas mais frias, sendo um indicador de atividade convectiva. A Figura 5.2 apresenta a imagem do aplicativo ForTraCC (Previsão a Curto Prazo da Evolução de Sistemas Convectivos), desenvolvido pelo DSA/CPTEC, sendo o fundo mais escuro correspondente à atividade convectiva. A Figura 5.3 corresponde à imagem do radar meteorológico de São Roque (CAPPI 3 km), operado pelo DECEA (em azul, verde e amarelo). Na Figura 5.4, temos a imagem da precipitação instantânea (em azul) dada pelo software Hidroestimador, que utiliza imagens do satélite GOES na banda infravermelha (canal 4). Nas Figuras 5.2 a 5.4, os pequenos círculos verdes, amarelos e vermelhos correspondem à chuva acumulada, medida por pluviômetros de estações meteorológicas do INMET.

Essa comparação mostra quais destas visualizações são mais convenientes para o monitoramento da atividade convectiva, assumindo que as imagens de radar sejam a forma mais precisa para essa finalidade. Uma vez que a cobertura dos radares meteorológicos é muito pequena no território brasileiro, as demais visualizações são desejáveis para tal monitoramento.

Adotando-se então a imagem de radar como referência para visualizar/monitorar a atividade convectiva, pode-se concluir que a imagem da densidade de descargas seja a que melhor se aproxima desta. Isso demonstra o potencial que o software EDDA tem para esse monitoramento, dada a pequena cobertura de radares meteorológicos no Brasil. O uso do software EDDA no CEMADEN parece demonstrar esse potencial. No caso, a grade utilizada abrange as longitudes de 36,5° a 62,5° Oeste e as latitudes 6,0° a 35,0° Sul, com resolução de 0,01°, num total de 7.545.501 pontos de grade e integrando descargas NS a cada 15 min para gerar os correspondentes campos de densidade de ocorrência. Considerando-se estes parâmetros, o software EDDA gera o campo de densidade em aproximadamente 1 s num processador corrente como, por exemplo, um Intel Xeon de 2,93 GHz, considerando-se a versão compilada com

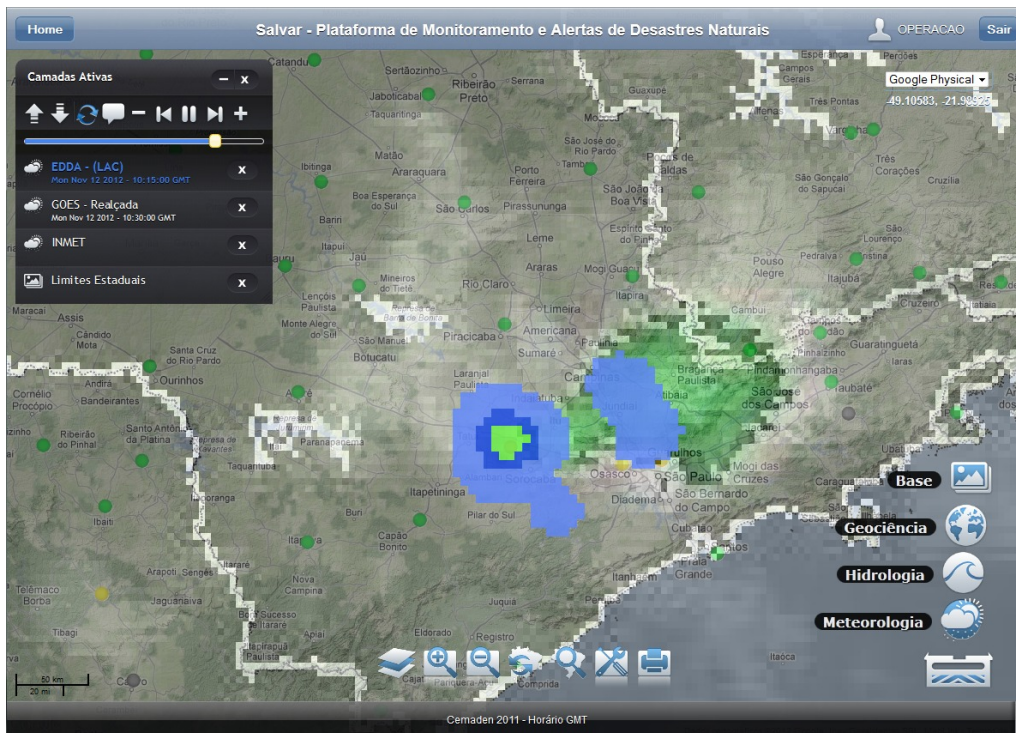


Figura 5.1 - Imagem de densidade de descargas (EDDA) sobreposta à imagem GOES reálçada.

o compilador da suíte Intel, não havendo, portanto, necessidade de paralelização. Para o mesmo caso, a versão compilada com o compilador GNU gcc, demandou um tempo de processamento de aproximadamente 1,5 s.

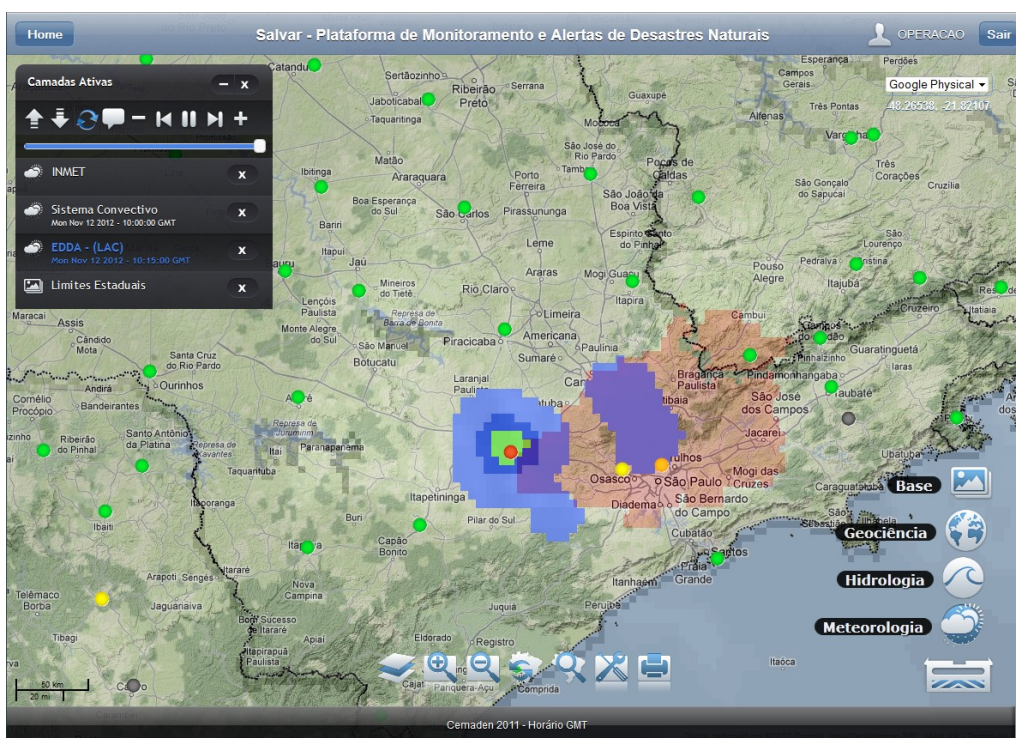


Figura 5.2 - Imagem de densidade de descargas (EDDA) sobreposta aos sistemas convectivos (ForTraCC) e às estações do INMET.



Figura 5.3 - Imagem de densidade de descargas (EDDA) sobreposta à imagem do radar de São Roque (CAPPi 3 km) e às estações do INMET.

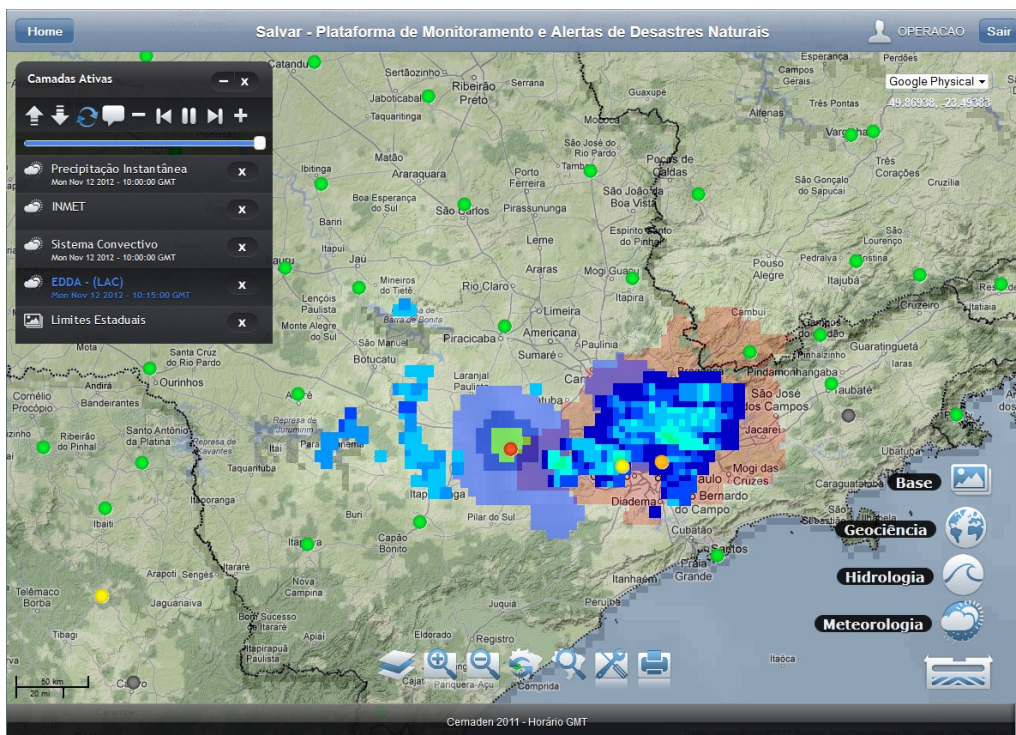


Figura 5.4 - Imagem de densidade de descargas (EDDA) sobreposta à precipitação instantânea estimada por satélite (Hidroestimador), aos sistemas convectivos e às estações INMET.

5.2 Uso da ferramenta EDDA com agrupamento de descargas para monitoramento e rastreamento de atividade convectiva

Apresentam-se a seguir os resultados obtidos no uso do software EDDA com agrupamento de descargas para o monitoramento de atividade convectiva relativa a duas tempestades que foram selecionadas como estudo de caso (STRAUSS et al., 2013). Adotaram-se aqui para a janela deslizante temporal a duração de 10 min e o avanço de 5 min, em função dos dados de descargas, conforme explicado na Seção 2.1.

A Seção 5.2.1 faz uma análise sinótica das tempestades selecionadas como estudo de caso para as seções subsequentes. A Seção 5.2.2 apresenta as imagens geradas pelo software para as tempestades selecionadas, comparando imagens de radar meteorológico e de células convectivas. A Seção 5.2.3 apresenta uma análise quantitativa da correlação entre estruturas de precipitação e células durante o período considerado. A Seção 5.2.4 mostra as trajetórias dos centroides de algumas células selecionadas. Finalmente, a Seção 5.2.5 mostra a evolução temporal de uma célula selecionada de cada evento.

5.2.1 Análise sinótica das tempestades selecionadas

Com o objetivo de investigar a correlação entre células eletricamente ativas e as estruturas de precipitação convectiva, dois eventos convectivos foram selecionados na proximidades de radares meteorológicos no estado de São Paulo durante a estação chuvosa (RICKENBACH et al., 2011). O evento A ocorreu em 16/01/2010 de 18:30 a 20:30 UTC e o evento B ocorreu em 19/01/2010 de 22:00 a 23:30 UTC.

O evento convectivo A causou forte chuva nas cidades de Bauru, Ribeirão Preto e Guarulhos em 16/01/2010. A Figura 5.5 (superior) mostra a imagem de satélite (à esquerda) correspondente do GOES-12 na banda do infravermelho.

O GOES é um satélite geostacionário, de 5 canais espectrais sendo um visível (0,55-0,75 μm), três canais infravermelhos (3,8-4,0 μm , 10,2-11,2 μm , 11,5-12,5 μm) e o canal de vapor d'água (6,5-7,0 μm). No canal visível, a resolução é 1 km. Nos canais infravermelhos, a resolução é de 4 km. No canal vapor d'água, a resolução é de 8 km.

Uma análise sinótica do evento A foi realizada a partir das análises do modelo atmosférico de circulação geral (AGCM) do CPTEC/INPE, que é usado para gerar as previsões de tempo e clima. Neste exemplo, utilizou-se a previsão sinótica de curto prazo (até 24 horas) para análise sinótica A previsão específica para 16/01/2010 às 18:00 UTC é mostrada na Figura 5.5 (inferior, à esquerda). Tal análise mostrou uma frente fria já em processo de frontólise, posicionada na latitude de $\sim 23^\circ\text{S}$) na costa brasileira com uma orientação NO-SE. Acima de Mar del Plata (latitude $\sim 34^\circ\text{S}$), existia um vórtice ciclônico em 500 hPa, e associado a este sistema, uma banda de nuvens acima do Sul do Brasil. No nível 250 hPa observava-se a presença do anticlone com centro sobre a Bolívia com uma crista avançando na latitude $\sim 23^\circ\text{S}$ acima dos estados de São Paulo (SP), Rio de Janeiro (RJ) e Minas Gerais (MG). Estes anticiclones são típicos da América do Sul nas temporadas de verão, geralmente conhecidos como “Alta da Bolívia” (AB), sendo associado à liberação de calor latente pela convecção sobre a Bacia Amazônica e a fluxo de calor sensível do platô boliviano (PRAKKI et al., 1998). Além da AB, outros fenômenos meteorológicos estão associados a convecção, tais como frentes frias e complexos convectivos de mesoescala. (RICKENBACH et al., 2011; CAVALCANTI, 2012). A frente fria associada à AB e o aquecimento diurno podem gerar muitas células convectivas acima do estado de SP. No evento estudado, as áreas de convecção em níveis baixos tiveram um deslocamento de NO para SE, mas a convecção em níveis superiores seguiu a circulação de BH, começando às 16:00 UTC e durando até a 24:00 UTC.

O segundo evento convectivo (B) era composto de algumas células convectivas que foram geradas durante a tarde, sobre o interior do estado de São Paulo. A Figura 5.5 (superior) é a imagem do satélite GOES-12 (à direita) na banda do infravermelha. Uma análise sinótica deste evento foi realizada com base em previsões do modelo atmosférico de circulação geral (AGCM) executado no CPTEC/INPE. A previsão específica para 19/01/2010 às 18:00 UTC é mostrada na Figura 5.5 (inferior, à direita). Estas células foram acionadas por aquecimento diurno e também devido a um componente dinâmico gerado por um processo ciclônico, centralizado em um ponto sobre o Oceano Atlântico na costa do estado do Rio Grande do Sul (RS). Às 09:00 UTC (não mostrado), foi possível observar uma queda na altura geopotencial em 250 hPa. O centro de baixa pressão associado se formou ao longo da costa do RS em torno das 21:00 UTC e propagou-se na direção NO-SE. Um padrão de circulação ao longo desta direção prevaleceu desde 12:00 UTC no nível de 850 hPa, dirigindo todas as células convectivas sobre o estado de São Paulo nas horas seguintes. No nível 250 hPa, as linhas de contorno mostram um sistema de circulação associado a uma área de alta pressão de mesoescala no setor leste e um padrão de circulação S-SO no sector oeste, de 12:00 a 24:00 UTC. A Figura 5.5, abaixo à direita, mostra esse sistema às 18:00 UTC. Essa área de alta pressão de mesoescala esteve claramente associada ao complexo convectivo de mesoescala (CCM) que se desenvolveu sobre o litoral do estado de São Paulo, provocando fortes chuvas sobre o Sul da cidade de São Paulo e ao longo de zonas litorâneas adjacentes. Por volta das 24:00 UTC, outro CCM pode ser observado no meio-oeste do estado, enquanto que o CCM anterior tinha perdido força, gerando basicamente precipitação estratiforme.

A fim de obter as taxas média e total de precipitação, duas áreas quadradas com lados de aproximadamente 55 km (0.5°) foram escolhidas, uma para o evento A, chamado Q_A , centrada no ponto com longitude $48^\circ 30' W$ e latitude $22^\circ 30' S$ e outra para o evento B, chamado Q_B , centrada no ponto com longitude $46^\circ 12' W$ e latitude $24^\circ S$. A taxa de precipitação média e a precipitação total foram calculados para cada área, considerando apenas “pixels molhados” ao longo da duração de cada evento. Taxas de precipitação média para os eventos A e B foram 12,3 e 11,8 mm/h, enquanto os totais de precipitação foram 24,5 e 17,7 mm, respectivamente.

5.2.2 Análise da correlação entre descargas e atividade convectiva para os eventos selecionados

A refletividade de radar em dBZ para o evento A em 16/01/2010 às 20:10 UTC é mostrada na parte superior da Figura 5.6. Os círculos correspondem aos intervalos

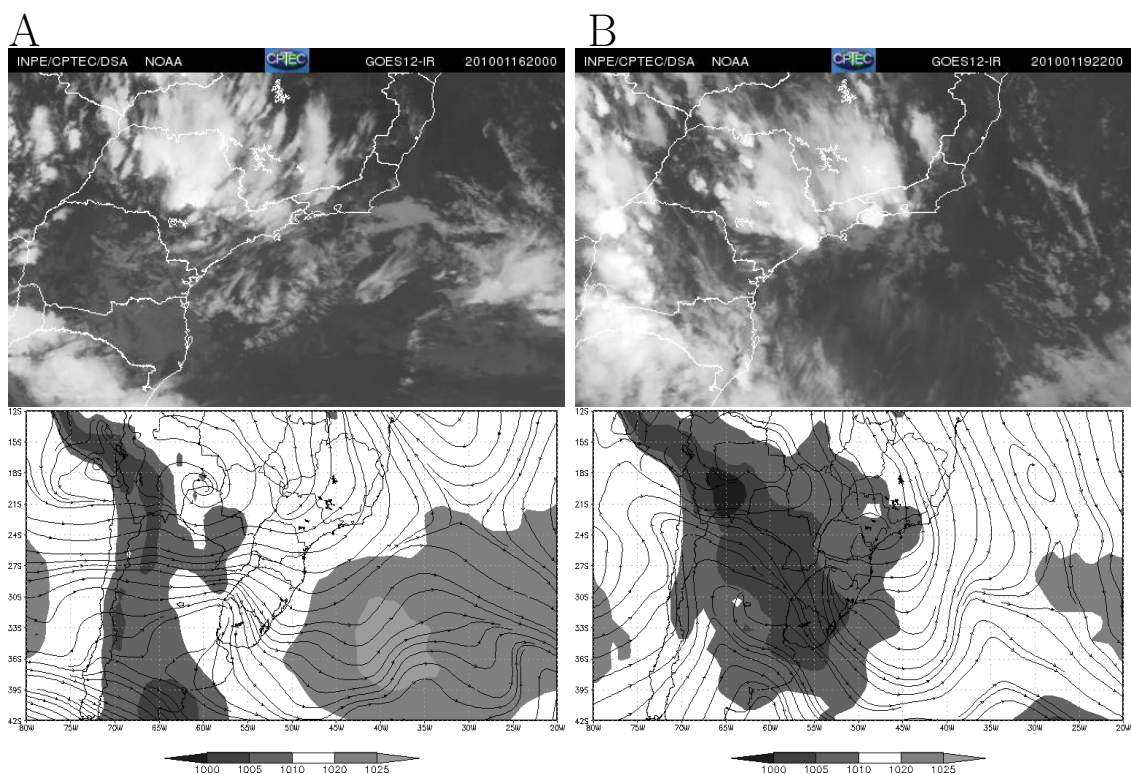


Figura 5.5 - (Superior) Imagens do satélite GOES-12 no canal infravermelho (canal 4 = $3,8-4,0 \mu\text{m}$), em 16/01/2010 às 20:00 UTC para o evento A (à esquerda) e em 19/01/2010 às 22:00 UTC para evento B (à direita). (Inferior) Previsão do modelo atmosférico de circulação geral (AGCM), do CPTEC para o evento A em 16/01/2010 18:00 UTC (à esquerda) e para o evento B em 19/01/2010 18:00 UTC (à direita) mostrando pressão na superfície (sombreada) e o padrão de circulação em 250 hPa (linhas de contorno).

de 150 km dos radares meteorológicos de Bauru (superior esquerdo) e São Roque (inferior direito). As partes central e inferior da Figura 5.6 mostram as mesmas estruturas de precipitação classificadas como convectiva (em preto) ou estratiforme (em cinza) de acordo com os critérios de Steiner. As duas figuras mostram células eletricamente ativas sobrepostas com as estruturas de precipitação. No entanto, a Figura 5.6 (centro) apresenta os limites das células eletricamente ativas obtidas exclusivamente pela estimacão de densidade. Os parâmetros empregados nesta estimacão foram: resolução de grade de $0,02^\circ$, largura de janela de $0,08^\circ$ e intervalo de tempo de 10 min. Os contornos das células correspondem a um limiar de densidade de $1 / (\text{grau}^2 \times \text{min})$. Por outro lado, a Figura 5.6 (inferior) mostra as células obtidas pelo agrupamento espaço-temporal proposto com uma janela deslizante de 10 min de duração e tamanho de passo de 5 min. Estes tempos são compatíveis com os dados de radar que foram interpolados para intervalos de 10 min. A estimacão de densidade então foi aplicada a cada célula do evento A e o limiar de densidade

a mesma foi aplicado para delinear os contornos de cada célula. Além disso, outro limiar mínimo de 3 descargas por célula foi aplicado, já que o agrupamento permite monitorar o número de descargas por célula.

Da mesma forma, refletividade de radar em dBZ para evento B em 19/01/2010 às 23:00 UTC é mostrada na parte superior da Figura 5.7, a parte central apresenta os limites das células eletricamente ativas obtidas por estimação de densidade unicamente, usando os mesmos parâmetros do evento A, e a figura inferior mostra as células obtidas pela abordagem proposta, também usando os mesmos parâmetros quanto ao evento.

O uso combinado de agrupamentos e da estimação da densidade de kernel permite identificar e visualizar mais precisamente a célula eletricamente ativa. Esta melhoria pode ser observada, comparando figuras do meio e inferior das Figuras 5.6 e 5.7 para os eventos A e B respectivamente. O uso de estimação de densidade sem agrupamento tende a aumentar indevidamente o número de células identificadas como apresentando precipitação convectiva, aumentando assim o número de “falsos positivos”. Por outro lado, a abordagem proposta permite visualizar as células, usando o mesmo limite de densidade, mas outro limiar, o número de descargas, pode ser aplicado a fim de descartar células com poucas descargas. Isso é possível, já que cada célula foi encontrada por agrupamento e que parâmetros como o número de descargas da célula ao longo do tempo são preservados.

5.2.3 Análise de correlação entre células eletricamente ativas e células de chuva para os eventos selecionados

A correlação entre as células eletricamente ativas do evento A e a chuva convectiva observada pelo radar às 20:10 UTC foi medida identificando-se os pixels de chuva que se sobrepõem aos pixels das células para a área considerada. Integra-se a densidade normalizada de ocorrência de descargas para esses pixels, obtendo-se um valor no intervalo $[0;1]$ (se estes pixels abrangessem a área inteira, esta densidade integrada seria 1). Essa densidade integrada é adotada como medida de correlação. Consideram-se aqui apenas as células com mais de 10 descargas no intervalo de 10 min centrado em 23:00 UTC.

No cálculo da correlação para ambos os eventos selecionados, adotaram-se três casas decimais, uma vez que esse foi o número de casas adotado para a densidade de ocorrência de descargas para fins de visualização.

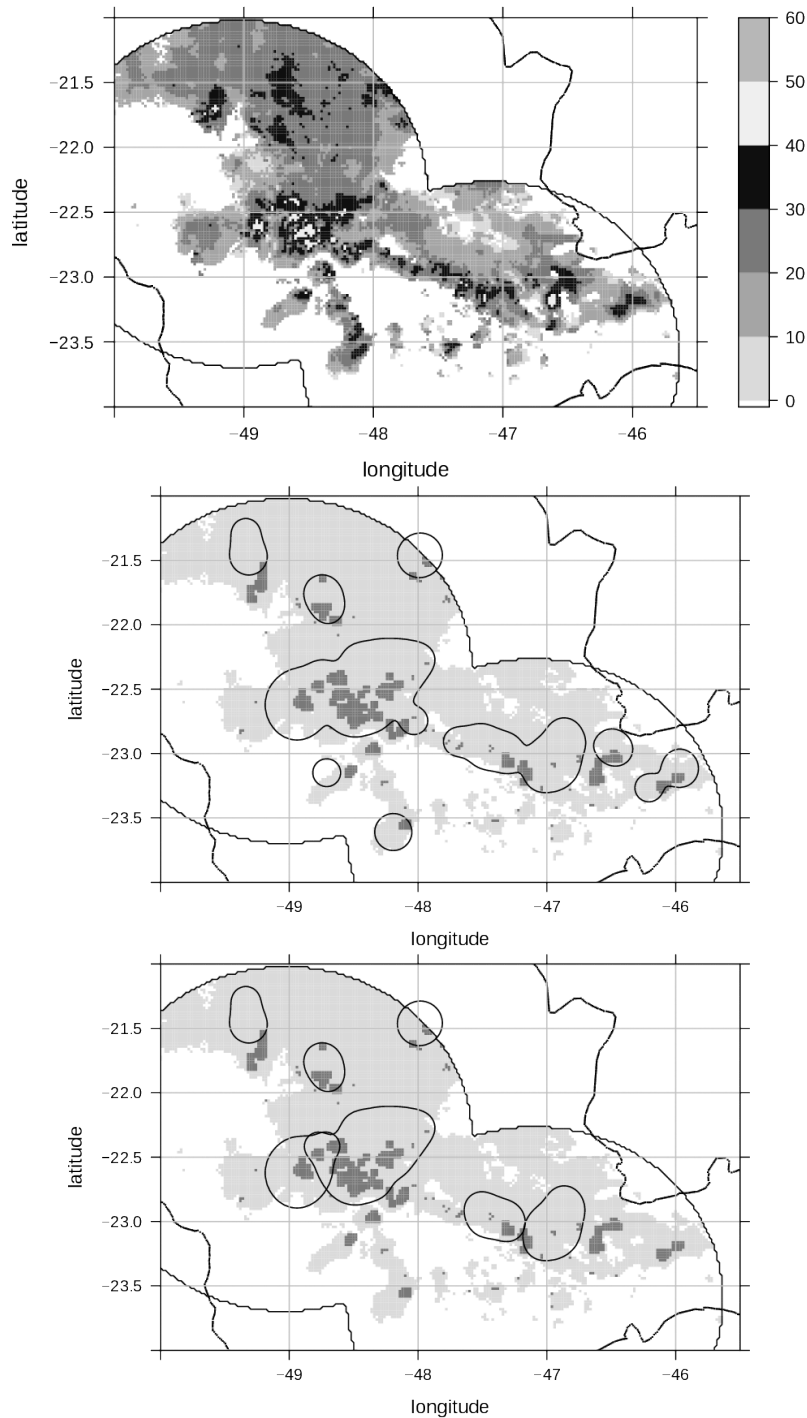


Figura 5.6 - Topo: Refletividade do radar (dBZ) do evento A em 16/01/2010 às 20:10 UTC. Meio e inferior: células eletricamente ativas do evento A detectado unicamente pela estimação de densidade (meio) e com o agrupamento (parte inferior). Tons de cinza correspondem a precipitação estratiforme, e tons escuros, a precipitação convectiva. Cada célula é mostrada por meio de seus contornos em um limiar de densidade determinado.

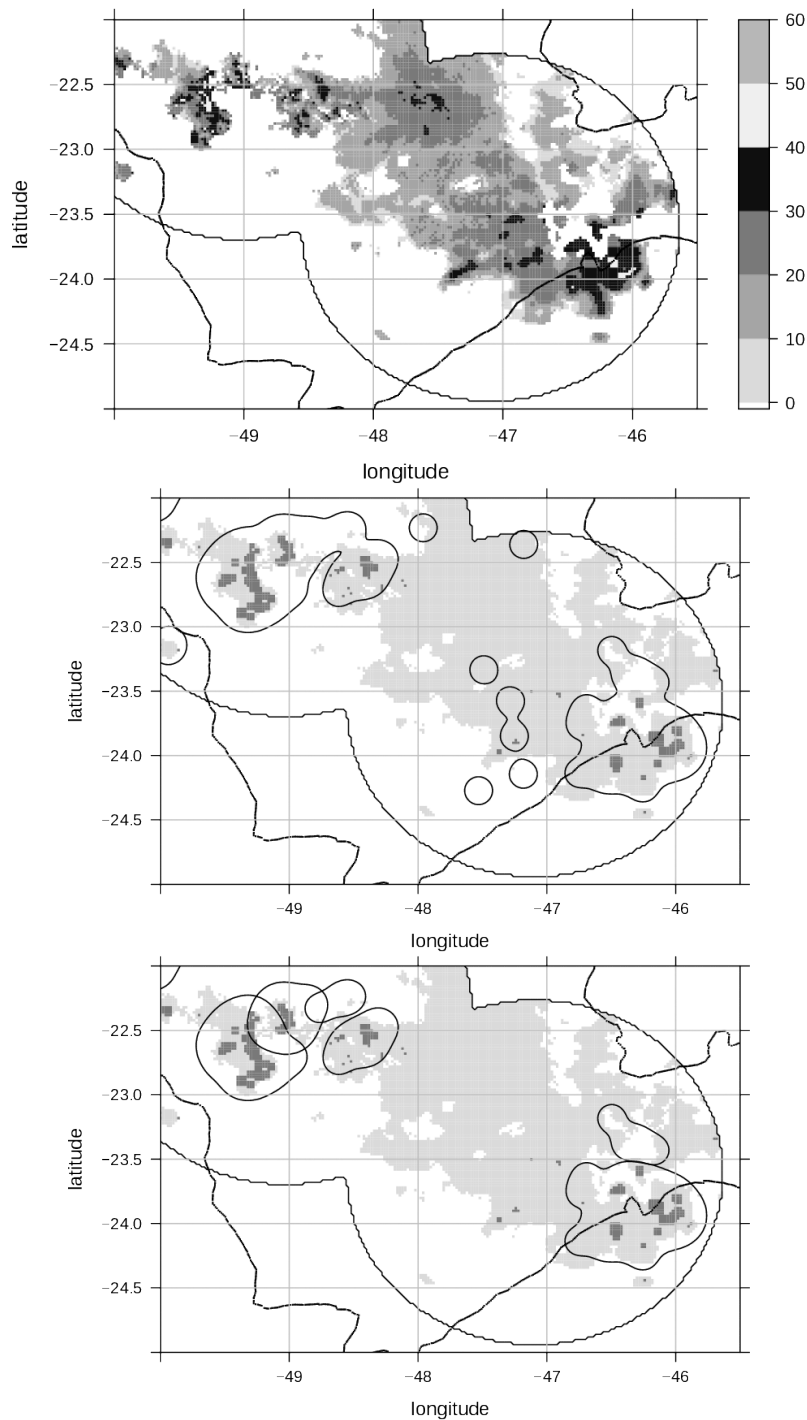


Figura 5.7 - Topo: Refletividade do radar (dBZ) do evento B em 19/01/2010 às 23:00 UTC. Meio e inferior: células eletricamente ativas do evento B detectado unicamente pela estimação de densidade (meio) e com o agrupamento (inferior). Tons de cinza correspondem a precipitação estratiforme e tons escuros, a precipitação convectiva. Cada célula é mostrada por meio de seus contornos em um limiar de densidade determinado.

Tabela 5.1 - Média ponderada das correlações entre as 3 células eletricamente ativas do evento A às 20:10 UTC e a chuva convectiva correspondente para diferentes valores de deslocamento (zero, até 1 pixel, ou até 3 pixels). Os três valores referem-se à correlação entre as células e a chuva observada na iteração anterior (20:00 UTC), a corrente (20:10 UTC) e o subsequente (20:20 UTC), respectivamente, para cada célula. As melhores correlações são destacadas.

ID	N	deslocamento		
		zero	até 1 pixel	até 3 pixels
# 1	66	0,623 /0,532/0,417	0,637 /0,555/0,453	0,637 /0,569/0,480
# 2	57	0,785 /0,710/0,470	0,885 /0,807/0,624	0,885 /0,807/0,710
# 3	18	0,289 /0,198/0,116	0,343 /0,229/0,222	0,343 /0,229/0,222
teste		0,646 /0,561/0,400	0,700 /0,615/0,492	0,700 /0,622/0,540
referência		0,592 /0,496/0,343	0,592 /0,517/0,403	0,592 /0,517/0,409

No entanto, já que não se pode esperar uma correlação espacial completa devido à dinâmica do evento convectivo, cada célula eletricamente ativa é considerada em sua posição original, mas também deslocada em todas as direções (para cima, para baixo, à esquerda, à direita, ou na diagonal) por um ou até três pixels, procurando a melhor correspondência com a chuva convectiva. Uma vez que uma correlação temporal completa pode não ocorrer, um deslocamento temporal é também considerado, ou seja, para cada célula, a correlação com a chuva convectiva é verificada para as imagens de radar anteriores e subsequentes, considerando também o deslocamento espacial, conforme mostrado na Tabela 5.1, para as três células com mais de 10 descargas, identificadas pelos ID (*identifier*) #1, #2, #3, sendo N o número de descargas de cada uma destas células. Nesta tabela, “teste” refere-se aos valores da correlação obtidos pela abordagem proposta, enquanto “referência” refere-se aos valores de correlação obtidos unicamente com a estimativa de densidade. Na mesma tabela, três valores da correlação são também apresentados para diferentes deslocamentos espaciais (zero, até 1 pixel, ou até 3 pixels): o primeiro está relacionado com a imagem do radar anterior (20:00 UTC), o segundo para a atual imagem de radar (20:10 UTC) e o terceiro para o subsequente (20:20 UTC). A melhor correlação é mostrada realçada. A Figura 5.8 mostra algumas das células que aparecem na Tabela 5.1.

A correlação entre as células eletricamente ativas do evento B e a chuva convectiva às 23:00 UTC foi realizada da mesma forma como a do evento A e pode ser vista na Tabela 5.2. As cinco células com mais de 10 descargas são identificadas por #1 até #5. Da mesma forma, correlações “teste” e “referência” são apresentadas para o evento B. Três valores de correlação são também apresentados para diferentes

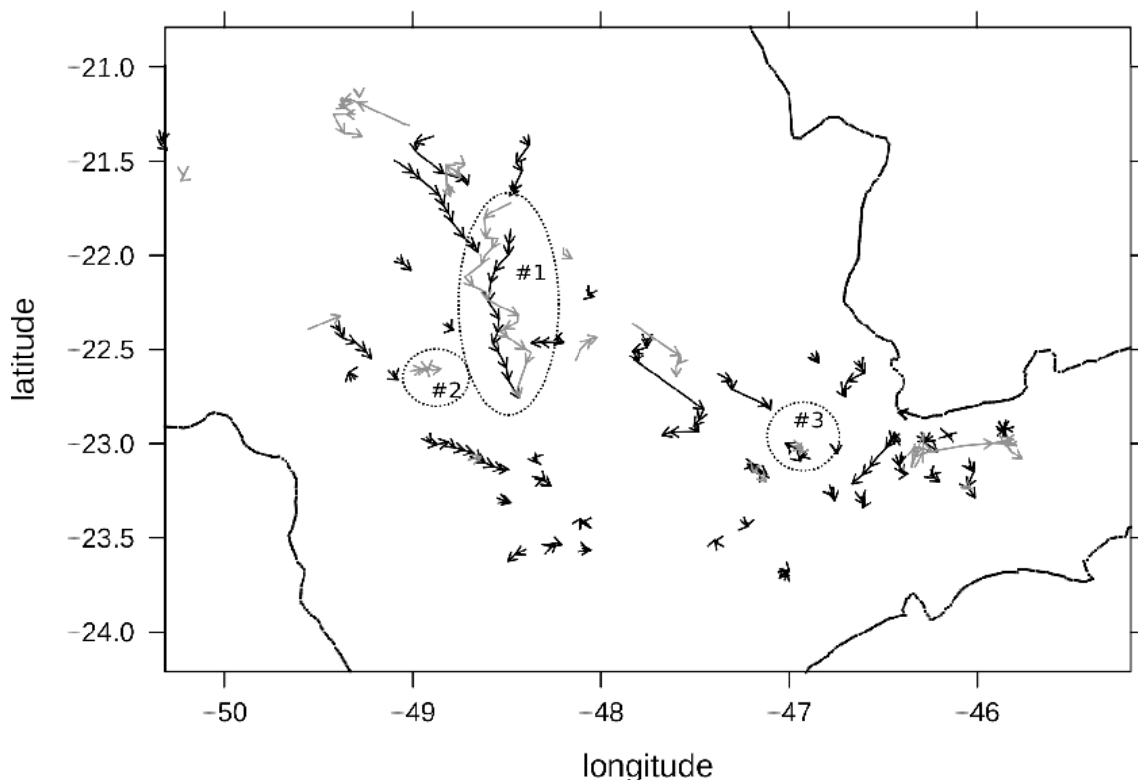


Figura 5.8 - Rastros no solo dos centroides das células eletricamente ativas (cinza) e células de precipitação (preto) no evento A (16/01/2010 de 18:30 para 20:30 UTC), destacando as células #1 para #3.

deslocamentos (zero, até 1 pixel ou até 3 pixels): o primeiro está relacionado com a imagem do radar anterior (22:50 UTC), o segundo com a atual (23:00 UTC) e a terceira com a subsequente (23:10 UTC). A melhor correlação é mostrada realçada. A Figura 5.9 mostra algumas das células que aparecem na Tabela 5.2.

Os resultados apresentados nessas tabelas para os eventos A e B mostram que a abordagem proposta (combinação de agrupamento e estimação de densidade) permite obter uma melhor correlação do que a aplicação de estimação de densidade unicamente.

Ao se considerar um deslocamento espacial, ou seja, uma vizinhança estendida, para a célula eletricamente ativa, isso pode ou não melhorar a correlação com a chuva observada, conforme essa vizinhança apresente ou não chuva, respectivamente. No caso negativo, a correlação se manterá igual mesmo com a vizinhança estendida. Isso explica a repetição de certos valores de correlação nas tabelas anteriores.

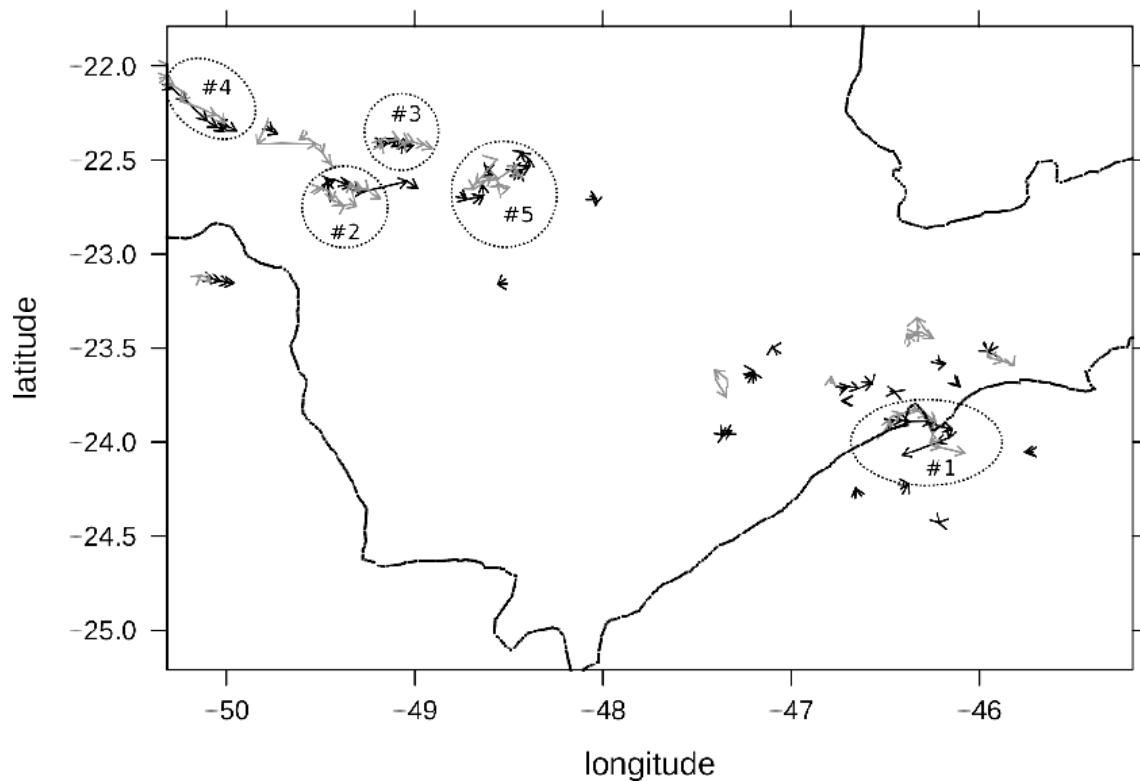


Figura 5.9 - Rastros no solo de células eletricamente ativas (cinza) e células de precipitação (preto) no evento B (19/01/2010 de 22:00 a 23:30 UTC), destacando as células #1 para #5.

Tabela 5.2 - Média ponderada das correlações entre as 5 células eletricamente ativas do evento B às 23:00 UTC e a correspondente chuva convectiva para diferentes valores de deslocamento espacial (nenhuma mudança, até 1 ou 3 pixels). Os três valores referem-se à correlação entre as células e a chuva observada na iteração anterior (22:50 UTC), a corrente (23:00 UTC) e o subsequente (23:10 UTC), respectivamente, para cada célula. As melhores correlações são destacadas.

ID	N	deslocamento		
		zero	até 1 pixel	até 3 pixels
# 1	199	0,387 /0,297/0,187	0,507 /0,326/0,209	0,535 /0,332/0,225
# 2	177	0,715 /0,497/0,239	0,731 /0,601/0,302	0,731 /0,663/0,396
# 3	76	0,749 /0,649/0,413	0,755 /0,723/0,577	0,755 /0,723/0,669
# 4	24	0,717 /0,377/0,050	0,717 /0,499/0,132	0,717 /0,540/0,262
# 5	12	0,589 /0,337/0,089	0,589 /0,462/0,122	0,589 /0,519/0,154
teste		0,584 /0,430/0,232	0,639 /0,499/0,294	0,651 /0,527/0,356
referência		0,563 /0,414/0,224	0,586 /0,470/0,275	0,586 /0,481/0,326

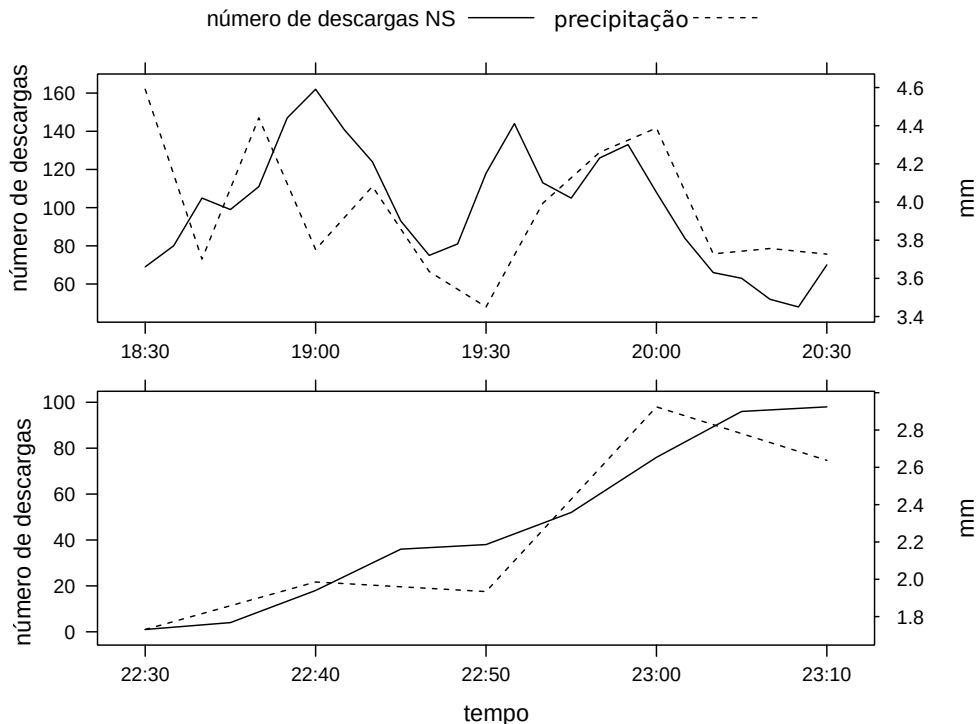


Figura 5.10 - Evolução temporal do número de descargas NS acumulado em 10 min e a precipitação média para a célula #1 do evento A (topo) e células #3 do evento B (inferior) que passou por cima de Bauru.

5.2.4 Comparação das trajetórias das células eletricamente ativas e das células de chuva convectiva para os eventos selecionados

Os centroides das células eletricamente ativas de ambos os eventos e os centroides das estruturas de precipitação correspondentes foram calculados para cada intervalo de 10 min.

Os rastros dos centroides das células #1 até #3 para o evento A são mostradas na Figura 5.8 (intervalo de 18:30 - 20:30 UTC), onde os rastros foram destacados por linhas de contorno. Analogamente, os rastros dos centroides das células #1 até #5 para o evento B (intervalo 21:00 - 23:30 UTC) são mostrados na Figura 5.9, também contornados. Em ambas as figuras, os rastros das célula com descargas aparecem em cinza e os de precipitação, em preto. Foram consideradas apenas células com mais de 10 descargas em qualquer intervalo de 10 min. Cada vez que a janela avança em 5 min, a célula é atualizada, e também o seu centroide. Os rastros são definidos pela sequência de posições dos centroides.

5.2.5 Considerações relativas à evolução temporal para os eventos selecionados

Pode-se observar na Tabela 5.1 (evento A) e a na Tabela 5.2 (evento B) que a melhor correlação entre células eletricamente ativas e precipitação é alcançada quando a imagem de radar anterior (10 min mais cedo) é considerada. Isto pode não ser um padrão consistente, dependendo dos eventos específicos. Por exemplo, a evolução temporal do número de descargas NS e da precipitação média (em mm), ambos acumulados em intervalos de 10 min, foram analisados para duas células particulares. A Figura 5.10 (superior) refere-se à evolução temporal da célula #1 do evento A, enquanto a Figura 5.10 (inferior) refere-se à célula #3 do evento B. Pode-se observar que as curvas apresentam padrões diferentes. Uma comparação extensiva da evolução temporal da precipitação e de descargas está fora do escopo deste trabalho, mas parece razoável supor que uma correlação temporal existe dentro de um certo intervalo de tempo e, por conseguinte, não iria interferir na abordagem proposta.

5.3 O software EPPA para previsão de ocorrência de precipitação

O software EPPA objetiva realizar a previsão da ocorrência e distribuição de chuva. Inicialmente, o software foi testado quanto à previsão da ocorrência de chuva (convectiva ou não) de qualquer intensidade, utilizando dados do modelo ETA 20 km. Os resultados foram satisfatórios. A seguir, passou-se à previsão específica da ocorrência de chuva forte e convectiva, ou seja, que verifique uma ou ambas condições, “forte”, acima de 7,5 mm/h e “convectiva” de acordo com o critério de Steiner, ainda utilizando dados do ETA 20 km. Todos os resultados obtidos com dados do ETA 20 km aparecem na Seção 5.3.1. O passo seguinte foi a previsão específica da ocorrência de chuva forte e convectiva com dados do ETA 5 km, conforme exposto na Seção 5.3.2. Esta última seção também inclui uma discussão específica sobre os parâmetros e esquemas referentes à otimização da árvore de decisão.

Nas seções seguintes, os resultados referem-se à previsão de ocorrência de chuva, por meio de visualizações, ou então, expressos pelo desempenho de classificação (índices POD e FAR). Este desempenho é relativo à habilidade do software EPPA em classificar corretamente um pixel como apresentando ou não ocorrência de chuva a partir de dados de análise e previsão do modelo ETA. O software EPPA é baseado em árvore de decisão, gerando-se árvores específicas para cada previsão de chuva desejada. De maneira geral, a árvore de decisão é treinada conforme descrito anteriormente na Seção 3.2.2, usando um conjunto de dados de treinamento, sendo a previsão de ocorrência gerada a partir do conjunto de dados de teste. Em alguns

casos, utilizou-se também um conjunto de dados de validação para otimizar a árvore. Repete-se abaixo a definição dos índices de desempenho POD e FAR:

$$\text{Probabilidade de detecção (POD)} = \text{VP} / (\text{VP} + \text{FN}) \times 100\% \quad (5.1)$$

$$\text{Taxa de falso alarme (FAR)} = \text{FP} / (\text{VP} + \text{FP}) \times 100\% \quad (5.2)$$

5.3.1 Previsão de ocorrência de precipitação com dados do modelo ETA 20 km

Nesta seção apresentam-se os resultados para a previsão da ocorrência de chuva (convectiva ou não) de qualquer intensidade. Esses resultados foram obtidos utilizando-se dados do modelo ETA 20 km do período de 1 a 26 de janeiro de 2010, considerando-se uma área coberta pelos radares de Bauru, Presidente Prudente e São Roque. Essas previsões foram comparadas às aquelas obtidas pela Ferramenta Objetiva de Previsão do Tempo do CPTEC/INPE. Os resultados referentes ao desempenho de classificação da Ferramenta Objetiva de Previsão do Tempo do CPTEC/INPE e do software EPPA são mostrados nas Tabelas 5.3 a 5.6 usando as métricas POD e FAR e considerando-se médias para cada quatro horários de previsão. Essas tabelas também mostram os resultados considerando-se as vizinhanças espaciais relaxadas de 3×3 pixels (60×60 km) e 5×5 pixels (100×100 km), considerando-se o ETA 20 km.

O desempenho médio da Ferramenta Objetiva de Previsão do Tempo do CPTEC/INPE é apresentado na Tabela 5.3, para pancadas de chuva com trovoadas (tipo 1) e na Tabela 5.4, para tempestades (tipo 2), sendo utilizadas análises (00 h) e previsões (06 a 90 h) do modelo ETA 20 km. Diferentemente do propósito original da ferramenta, projetada para auxílio à previsão de eventos os três tipos (sendo o tipo 3, granizo), o desempenho avaliado é referente à previsão de ocorrência de chuva (convectiva ou não, forte ou fraca), o que teoricamente favorece o desempenho da ferramenta. Observa-se nestas tabelas que o POD ficou muito abaixo de 50% na maioria dos casos, sendo que o FAR era frequentemente maior que o POD para o caso correspondente. É preciso fazer a ressalva que essa ferramenta é utilizada juntamente com a visualização de outros dados meteorológicos, de forma que, na prática, seu desempenho de classificação deve ser melhor.

Tabela 5.3 - Desempenho médio da Ferramenta Objetiva de Previsão do Tempo do CP-TEC/INPE para previsão de pancadas de chuva com trovoadas (tipo 1). O valores referem-se a cada conjunto de quatro previsões do ETA 20 km considerando diversas vizinhanças.

previsões (h)	vizinhança (pixels)					
	nula		3×3		5×5	
	POD	FAR	POD	FAR	POD	FAR
00 06 12 18	22,2	72,8	32,3	46,3	41,4	30,2
24 30 36 42	20,5	67,9	31,0	37,0	38,8	20,0
48 54 60 66	20,3	74,6	31,9	49,0	41,2	33,6
72 78 84 90	22,8	72,1	33,9	44,0	42,5	26,9

Tabela 5.4 - Desempenho médio da Ferramenta Objetiva de Previsão do Tempo do CP-TEC/INPE para previsão de tempestades (tipo 2). O valores referem-se a cada conjunto de quatro previsões do ETA 20 km considerando diversas vizinhanças.

previsões (h)	vizinhança (pixels)					
	nula		3×3		5×5	
	POD	FAR	POD	FAR	POD	FAR
00 06 12 18	1,61	26,1	2,76	36,2	4,07	21,5
24 30 36 42	2,25	27,5	5,16	33,3	8,23	16,7
48 54 60 66	1,96	16,7	4,70	50,4	7,41	30,1
72 78 84 90	4,00	24,2	7,69	43,7	11,44	24,5

Tabela 5.5 - Desempenho médio da árvore de decisão (EPPA) para previsão de ocorrência de precipitação, treinando e testando com dados de análise e previsão do modelo ETA 20 km.

previsões (h)	vizinhança (pixels)					
	nula		3×3		5×5	
	POD	FAR	POD	FAR	POD	FAR
00 06 12 18	47,7	68,5	85,4	44,6	94,3	30,4
24 30 36 42	45,0	68,1	83,4	42,4	92,6	27,7
48 54 60 66	47,4	66,6	82,7	40,6	93,7	26,1
72 78 84 90	44,8	70,3	80,8	43,7	92,7	25,5

Tabela 5.6 - Desempenho médio da árvore de decisão (EPPA) para previsão de ocorrência de precipitação, treinando com a análise (00 UTC) e previsões de 06 – 18 UTC, e testando com as demais previsões do modelo ETA 20 km.

previsões (h)	vizinhança (pixels)					
	nula		3×3		5×5	
	POD	FAR	POD	FAR	POD	FAR
24 30 36 42	42,0	30,0	80,9	44,4	93,1	28,1
48 54 60 66	45,8	29,0	82,4	46,7	92,4	31,8
72 78 84 90	42,9	28,6	81,6	47,0	93,5	31,6

As Tabelas 5.5 e 5.6 referem-se ao desempenho do software EPPA para previsão de ocorrência de chuva (convectiva ou não, forte ou fraca), usando dados do modelo ETA 20 km. Observa-se que o desempenho do EPPA sem vizinhança relaxada equipara-se ao desempenho da ferramenta com a maior vizinhança considerada, atingindo valores altos de POD e baixos de FAR. Esse desempenho obviamente melhora à medida que se consideram vizinhanças maiores. No tocante aos esquemas de treinamento, nos testes referentes à Tabela 5.5, para cada conjunto de quatro previsões, o conjunto de treinamento foi formado por 80% dos horários de previsões, escolhidas aleatoriamente, deixando o restante para testes. Isso caracteriza um esquema de treinamento com *hold out*. Isso significa considerar no treinamento todos os pixels para cada uma dessas previsões que compõem o conjunto de 80% e, analogamente, considerar todos os pixels dos 20% de previsões restante. Foram treinadas quatro árvores de decisão, cada uma correspondente a um dos quatro conjuntos de quatro previsões. Na Tabela 5.6, é apresentado o resultado de se treinar uma única árvore com o conjunto formado pelas análises das 00 h e apenas com as previsões de 06, 12 e 18 h, enquanto que o teste usou dados das previsões de 24 a 90 h, sendo o desempenho avaliado para os três conjuntos de quatro previsões subsequentes.

A análise dos resultados apresentados nas Tabelas 5.3 a 5.6 mostra que a Ferramenta Objetiva de Previsão do Tempo do CPTEC/INPE apresentou baixo desempenho de classificação, mesmo considerando-se a vizinhança relaxada de 5×5 pixels. Por outro lado, o software EPPA apresentou um desempenho de classificação aceitável para vizinhança relaxada de 3×3 pixels, e melhor ainda para vizinhança de 5×5 pixels.

A seguir, comparou-se o desempenho de classificação da Ferramenta Objetiva de Previsão do Tempo do CPTEC/INPE e do software EPPA com dados do modelo ETA 20 km para três tempestades selecionadas no mês de janeiro de 2010. Para estas tempestades, as imagens correspondentes do satélite GOES-12 na banda infravermelha, dos campos de pressão e das linhas de corrente e da refletividade medida por radar são apresentadas na Figura 5.11. A primeira tempestade ocorreu no dia 14/01/2010 e foi analisada para o horário de 18:00 UTC, onde se observa uma grande área de instabilidade sobre os estados do PR, SP, RJ. Esta instabilidade foi organizada por um sistema frontal atuante sobre o oceano próximo ao litoral sul do Brasil e por uma crista sobre o estado de SP, que organizou a convecção sobre o norte de SP e sudoeste e sul de MG. A segunda tempestade ocorreu no dia 20/01/2010, sendo analisada para o horário de 00 UTC, quando havia uma baixa fria em processo de ciclogênese sobre o Uruguai que organizava um canal de umidade sobre o interior do continente. Sobre o estado de SP havia uma crista associada à Alta da Bolívia

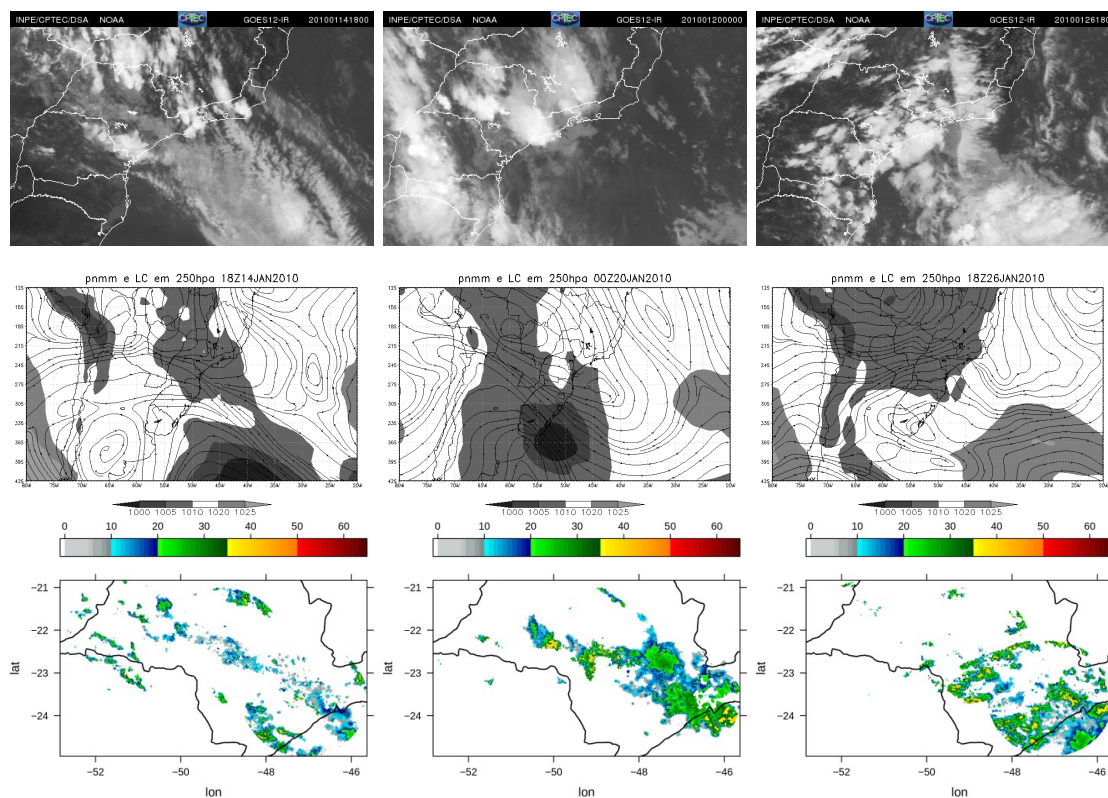


Figura 5.11 - Imagens meteorológicas relativas às três tempestades consideradas: evento de 14/01/2010 às 18:00 UTC (esquerda), evento de 20/01/2010 às 00:00 UTC (centro) e evento de 26/01/2010 às 18:00 UTC (direita). Na fileira superior aparecem as imagens do satélite GOES-12 na banda infravermelha, na fileira do meio os correspondentes campos de pressão e linhas de corrente (LC) e, na fileira inferior, as refletividades (dBZ) medidos pelos radares meteorológicos.

e que favorecia o desenvolvimento de células convectivas sobre todo o centro-sul do Brasil. A terceira tempestade foi no dia 26/01/2010, sendo analisada para o horário de 18:00 UTC, quando não havia qualquer sistema baroclínico atuante sobre o estado de SP. O desenvolvimento de células convectivas sobre o leste do estado de São Paulo se deu pela fraca divergência gerada pelo Jato Subtropical, cuja saída polar estava sobre este setor, e pelo aquecimento diurno.

Os resultados correspondentes a essas três tempestades aparecem nas Figuras 5.12, 5.13 e 5.14, e nas correspondentes Tabelas 5.7, 5.8 e 5.9. Foram usados dados de quatro diferentes previsões do modelo, feitas com antecedências diferentes e considerada sempre uma vizinhança relaxada de 3×3 pixels. Observa-se um desempenho satisfatório do software EPPA com dados do modelo ETA 20 km para previsão de ocorrência ou não de chuva. Nota-se também que, para a maioria das previsões, o desempenho do software EPPA foi melhor comparativamente à Ferramenta Objetiva

de Previsão do Tempo do CPTEC/INPE, considerando-se os pares de valores POD e FAR.

Tabela 5.7 - Comparação dos desempenhos da Ferramenta Objetiva de Previsão do Tempo do CPTEC/INPE e do software EPPA na previsão de ocorrência de precipitação para a tempestade do dia 14/01/2010 às 18:00 UTC para quatro previsões diferentes.

previsões (h)	Ferramenta Objetiva		software EPPA	
	POD	FAR	POD	FAR
00 + 18 h	34,3	11,7	51,4	21,9
24 + 18 h	43,1	7,5	66,9	23,7
48 + 18 h	11,1	11,1	63,0	15,0
72 + 18 h	16,6	11,4	70,2	17,9

Tabela 5.8 - Comparação dos desempenhos da Ferramenta Objetiva de previsão do Tempo do CPTEC/INPE e do software EPPA na previsão de ocorrência de precipitação para a tempestade do dia 20/01/2010 às 00:00 UTC para quatro previsões diferentes.

previsões (h)	Ferramenta Objetiva		EPPA	
	POD	FAR	POD	FAR
00 + 00 h	5,3	90,0	38,7	43,5
24 + 00 h	9,3	88,0	33,3	33,3
48 + 00 h	62,7	40,7	60,0	41,9
72 + 00 h	64,0	31,0	18,7	41,7

Finalmente, avaliou-se o desempenho de classificação do software EPPA com os mesmos dados do modelo ETA 20 km (mês de janeiro de 2010) para previsão da ocorrência e distribuição de chuva forte e convectiva, conforme aparece na Tabela 5.10. Observa-se que os valores de POD são ruins, abaixo dos correspondentes valores de FAR para o caso sem vizinhança. Além disso, para a vizinhança 3×3, embora os valores de POD sejam altos, os valores de FAR são também inaceitavelmente altos.

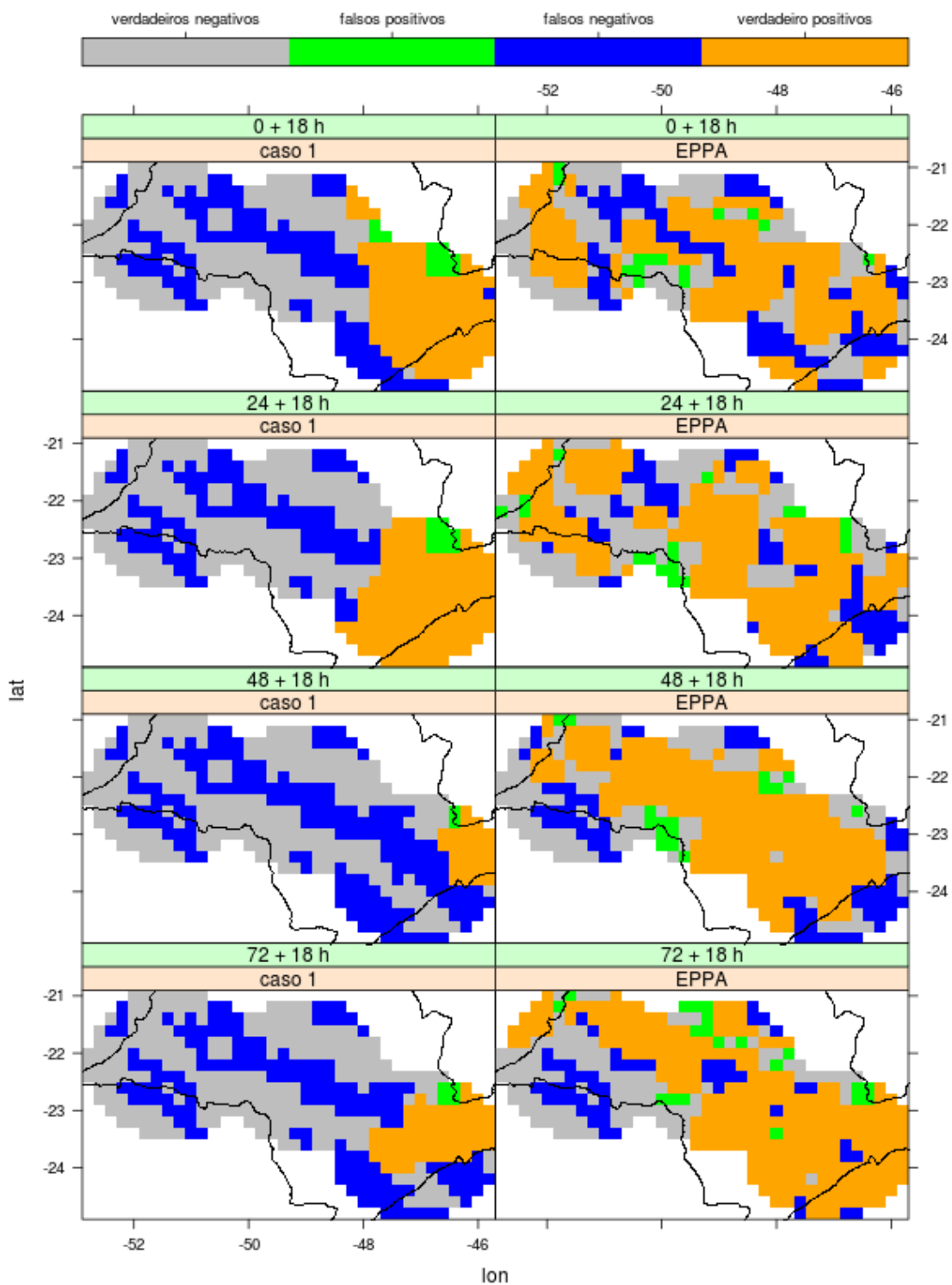


Figura 5.12 - Comparação das classificações efetuadas pela ferramenta de previsão objetiva do CPTEC (caso 1), à esquerda, e pelo software EPPA para previsão de ocorrência de precipitação, à direita, para a tempestade do dia 14/01/2010 às 18:00 UTC, considerando-se quatro diferentes previsões.

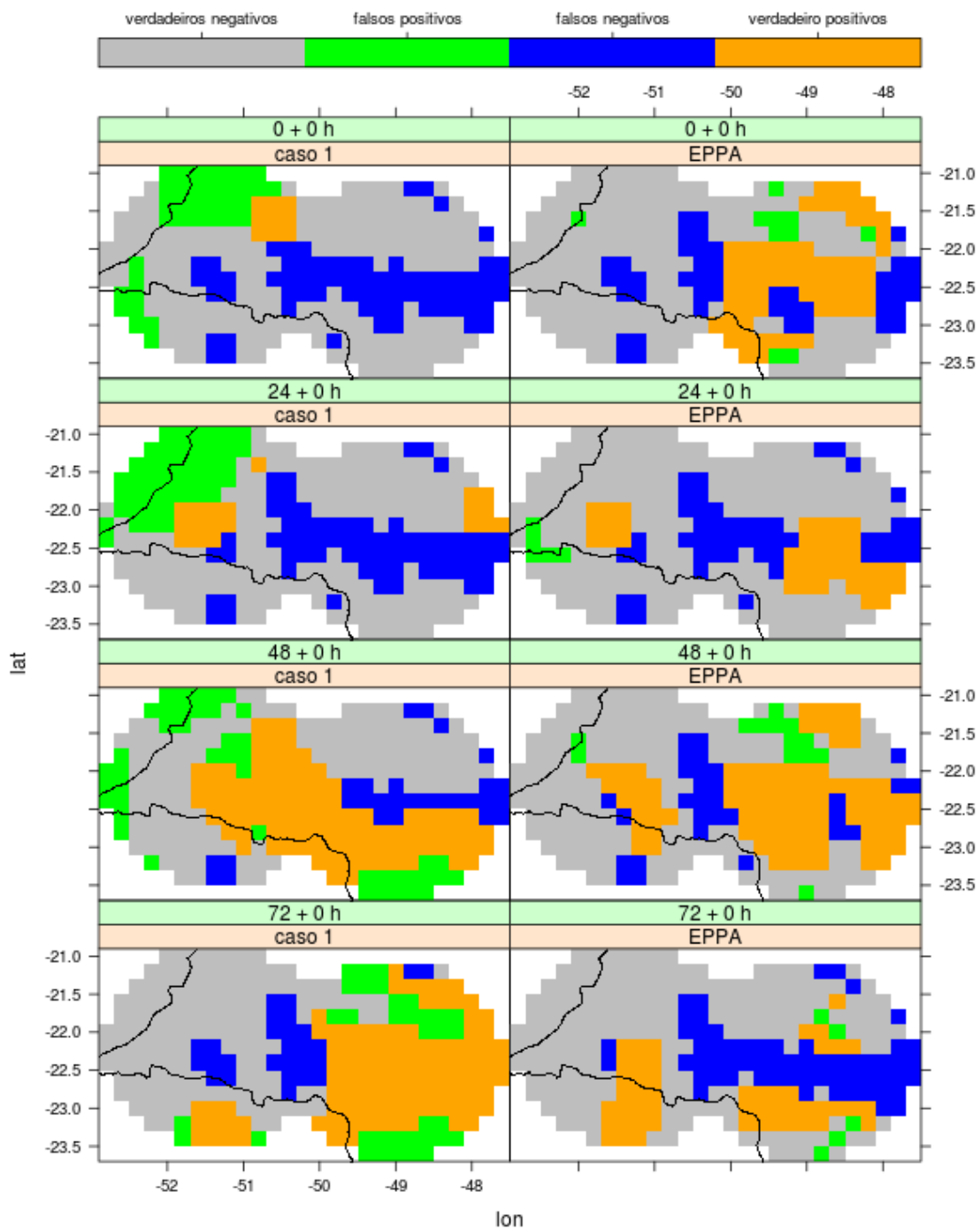


Figura 5.13 - Comparação das classificações efetuadas pela ferramenta de previsão objetiva do CPTEC (caso 1), à esquerda, e pelo software EPPA para previsão de ocorrência de precipitação, à direita, para a tempestade do dia 20/01/2010 às 00:00 UTC, considerando-se quatro diferentes previsões.

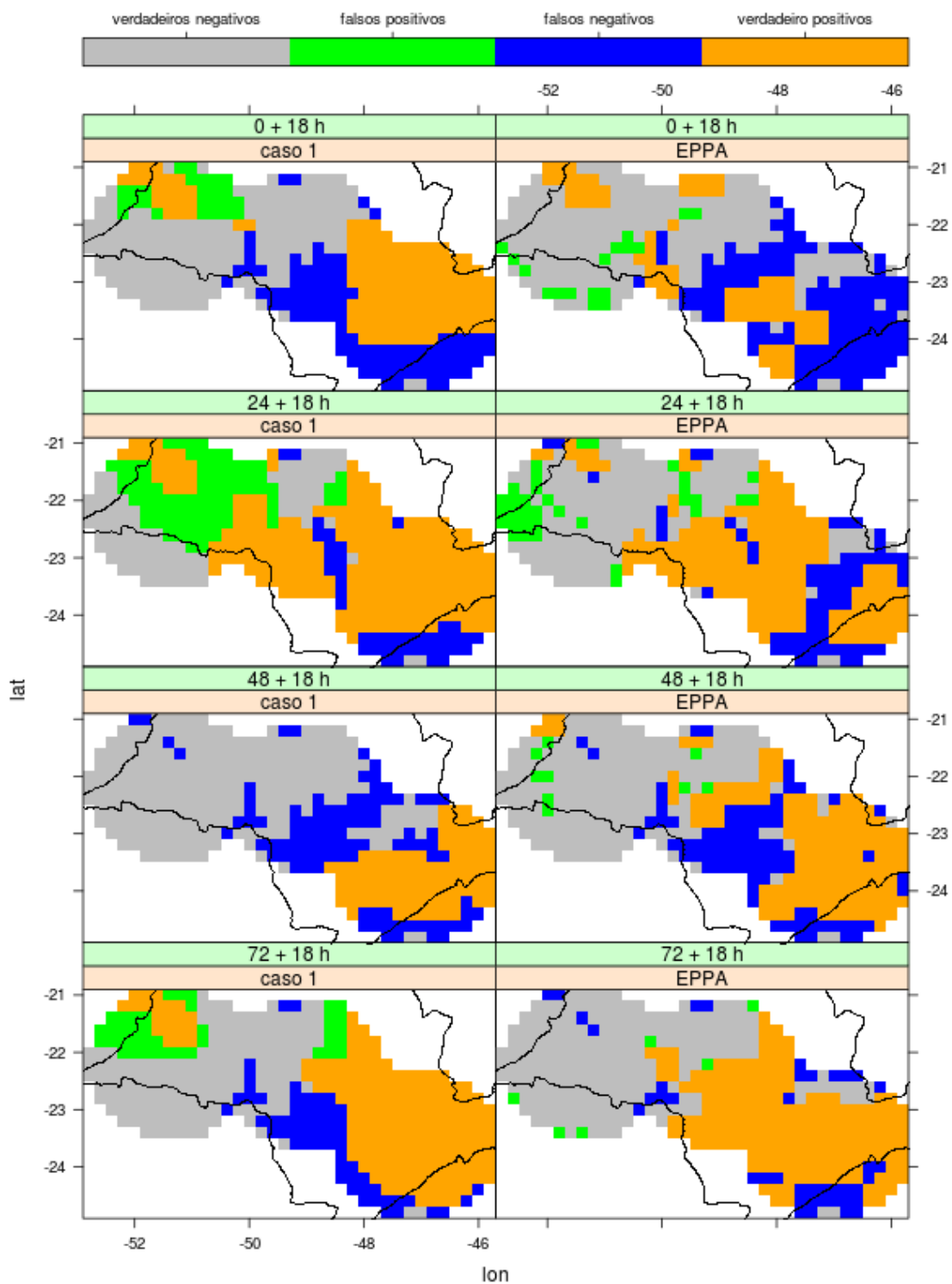


Figura 5.14 - Comparação das classificações efetuadas pela ferramenta de previsão objetiva do CPTEC (caso 1), à esquerda, e pelo software EPPA para previsão de ocorrência de precipitação, à direita, para a tempestade do dia 26/01/2010 às 18:00 UTC, considerando-se quatro diferentes previsões.

Tabela 5.9 - Comparação dos desempenhos da Ferramenta Objetiva de Previsão do Tempo do CPTEC/INPE e do software EPPA na previsão de ocorrência de precipitação para a tempestade do dia 26/01/2010 às 18:00 UTC para quatro previsões diferentes.

previsões (h)	Ferramenta Objetiva		EPPA	
	POD	FAR	POD	FAR
00 + 18 h	46,0	20,5	23,9	51,5
24 + 18 h	76,1	32,6	62,5	76,1
48 + 18 h	44,9	0,0	52,8	22,5
72 + 18 h	66,5	21,9	76,7	8,1

Tabela 5.10 - Desempenho médio da árvore de decisão (EPPA) para previsão de ocorrência de chuva forte ou convectiva, treinando com a análise (00 UTC) e previsões de 06 – 18 UTC, e testando com as demais previsões do modelo ETA 20 km.

previsões (h)	vizinhança (pixels)			
	nula		3×3	
	POD	FAR	POD	FAR
24 30 36 42	54,2	94,5	89,8	83,1
48 54 60 66	50,7	96,4	91,3	88,6
72 78 84 90	55,0	95,9	90,8	86,4

5.3.2 Previsão de ocorrência de precipitação com dados do modelo ETA 5 km

Nesta seção aparecem os resultados para a previsão da ocorrência de chuva forte e convectiva. Esses resultados foram obtidos utilizando-se dados do modelo ETA 5 km para os verões de 2009/2010 (período de 1 a 26 de janeiro de 2010) e 2010/2011 (meses de dezembro de 2010 e janeiro e fevereiro de 2011), considerando-se uma área coberta pelos radares de Bauru, Presidente Prudente e São Roque. Esta seção discute o balanço entre o número de falsos positivos (previsão de chuva forte e convectiva que não ocorreu) e o número de falsos negativos (chuva forte e convectiva que ocorreu, mas não foi prevista). Convém repetir que os falsos negativos são mais indesejáveis que os falsos positivos, uma vez que a não previsão de uma ocorrência de chuva forte ou convectiva é mais crítica que um falso alarme.

Os dados do modelo ETA 5 km foram divididos aleatoriamente em conjuntos de treinamento (70%), validação (15%) e teste (15%). O conjunto de validação foi usado para ajustar os parâmetros ou selecionar esquemas relativos aos dados e à árvore de decisão. Estes parâmetros e esquemas são discutidos a seguir:

- Proporção entre instâncias positivas e negativas nos dados: a proporção entre instâncias positivas (correspondentes a chuva forte ou convectiva) e negativas (chuva estratiforme ou fraca ou ausência de chuva) é de aproximadamente 1:30 no caso dos dados de janeiro de 2010, o que pode contribuir para que as instâncias positivas sejam mal classificadas. A técnica de alterar essa proporção pela diminuição das instâncias negativas e foi denotada aqui como “sub-amostragem”, sendo seu efeito piorar o POD (ou seja, diminuí-lo), enquanto melhora o FAR (ou seja, também diminuí-lo). Testou-se uma subamostragem até 1:25 com piora do desempenho de classificação e, conseqüentemente, preferiu-se manter a amostragem original de 1:30. No caso dos dados do verão de 2010/2011, essa proporção é de 1:4, sendo também mantida a amostragem original.
- Esquema de poda da árvore: a poda é feita com base numa medida de impureza, no caso o índice Gini. Entretanto, observou-se que a poda piora o desempenho de classificação, diminuindo os falsos positivos, mas aumentando muito os falsos negativos, estes mais indesejáveis que os primeiros, uma vez que o objetivo da classificação é detetar ocorrência de chuva forte e convectiva.

- Esquemas de *hold out*: estes esquemas relacionam-se com a seleção de dados para composição dos conjuntos de treinamento, validação e teste. Um primeiro esquema é treinar a árvore usando somente os dados das análises diárias das 0 UTC e validar/testar com os dados dessas análises e também das previsões diárias de 01 a 23 UTC. Um segundo esquema inclui também no conjunto de treinamento as análises e as previsões, mas separando aleatoriamente 70% dos dados para o treinamento, 15% para validação e 15% para teste, o que caracteriza um conjunto de dados de treinamento com *hold out* aleatório. Constatou-se que os melhores resultados foram obtidos com segundo esquema.
- Tamanho da vizinhança relaxada espacial: observa-se que um tamanho maior tende a aumentar o POD (desejavelmente), mas também aumentar o FAR (indesejavelmente).
- Vizinhança relaxada temporal: este esquema considera na classificação a chuva convectiva e forte ocorrida num intervalo de tempo estendido de 30 min antes até 30 min depois em relação à previsão do modelo. Essa vizinhança temporal foi adotada considerando-se a melhor resolução temporal do modelo. Analogamente à vizinhança relaxada espacial, sua adoção tende a aumentar o POD, mas também o FAR.
- Limiar de impureza: este limiar se aplica igualmente para cada nó da árvore; um limiar nulo ou baixo implicaria em baixo número de falsos positivos, mas num alto número de falsos negativos, o que seria indesejável; os testes realizados demonstraram que um limiar conveniente é 10%.
- Balanco entre falsos positivos (FP) e falsos negativos (FN): eventualmente, pode-se diminuir o número de FNs às custas de um aumento dos FPs, sendo que, conseqüentemente, o POD e o FAR aumentam. Esse esquema somente é razoável se o POD aumentar para valores altos, mas se o FAR aumentar para valores não muito superiores a 50%.

O desempenho de classificação obtido com o conjunto de validação indica que, em geral, os métodos de ajuste acima tendem a reduzir o número de falsos negativos, mas às custas de um maior número de falsos positivos. Assumiu-se que uma proporção desejável entre falsos negativos e falsos positivos seja de 1:2. A Figura 5.15 é o gráfico que ilustra o número de falsos negativos em função dos falsos positivos, de forma a avaliar a influência de alguns desses ajustes, no caso, o uso de poda e uso de

Tabela 5.11 - Desempenho médio da árvore de decisão (EPPA) para previsão de ocorrência de precipitação forte e convectiva, treinando e testando com dados de análise e previsão do modelo ETA 5 km, considerando-se ou não vizinhança temporal de ± 30 min e vizinhança espacial de 3×3 pixels.

Vizinhança	nenhuma	apenas temporal	apenas espacial	espacial e temporal
POD	32%	32%	75%	75%
FAR	60%	41%	34%	24%

sub-amostragem. Cada ponto da figura corresponde à proporção obtida entre falsos negativos e falsos positivos para uma dada configuração. Os dois segmentos de reta que partem da origem representam as proporções de 1:1 e 1:2 entre falsos negativos e falsos positivos. Essa figura é referente a uma vizinhança relaxada de 15×15 km (correspondente a um conjunto de 3×3 pixels do modelo ETA 5 km). Na mesma figura aparecem pares de pontos correspondentes à amostragem original (30:1) e à sub-amostragem 25:1 para diversos limiares de impureza, sendo que pontos mais à direita tem limiares de impureza mais altos. Observa-se que os pontos com sub-amostragem sempre estão ligeiramente abaixo e bastante à direita, ou seja, implicam em um número ligeiramente menor de falsos negativos, mas às custas de um número bem maior de falsos positivos. Nesse gráfico aparecem também dois pontos correspondentes ao uso de poda (sem sub-amostragem). Esses dois pontos são os melhores obtidos com uso de poda e mostram que seu desempenho de classificação é inaceitável, pois implicam num número muito alto de falsos negativos (POD baixo), embora minimizem os falsos positivos (FAR baixo). Assim, optou-se por não usar poda nem sub-amostragem, escolhendo-se um limiar de impureza próximo ao segmento de reta 1:2, que representa a proporção entre falsos negativos e falsos positivos.

A Tabela 5.11 refere-se ao desempenho do software EPPA para previsão de ocorrência de chuva forte e convectiva, usando dados do modelo ETA 5 km para o mês de janeiro de 2010 na área considerada (valores médios de POD e FAR). A árvore de decisão foi gerada considerando-se particionamento aleatório dos dados (análises e previsões até 24 h) em 70%, 15%, 15% para treinamento, validação e teste respectivamente. A árvore de decisão correspondente foi ajustada para o ponto mais próximo do segmento de reta 1:2 sem sub-amostragem da Figura 5.15. Os resultados consideram também vizinhança espacial de 3×3 pixels, vizinhança temporal de ± 30 min, ou ambas.

A previsão de chuva forte e convectiva a partir de dados do ETA 5 km foi também testada para as mesmas três tempestades selecionadas no mês de janeiro de 2010

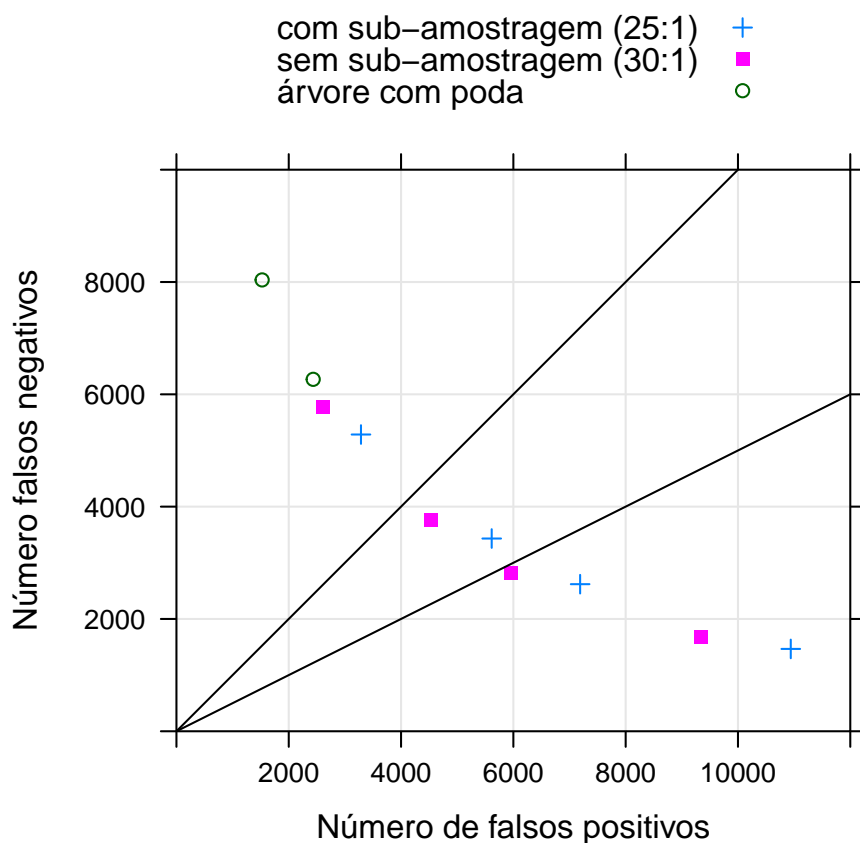


Figura 5.15 - Gráfico do número de falsos negativos em função dos falsos positivos para o conjunto de dados de validação do modelo ETA 5 km correspondente a janeiro de 2010, considerando-se uma vizinhança de 3×3 pixels. As ordenadas variam de 0 a 10000 e as abscissas, de 0 a 12000.

que foram descritas na seção anterior. Para estas tempestades foi gerada uma única árvore de decisão treinada com um conjunto de dados particionados aleatoriamente e com parâmetros similares às árvores de decisão referentes à Tabela 5.11, porém sem nenhuma vizinhança espacial ou temporal. Os resultados para a tempestade do dia 14/01/2010 nos horários 15, 18, e 21 UTC, e o horário 0 UTC do dia seguinte aparecem na Figura 5.16 e na Tabela 5.12. Os resultados para a tempestade do dia 19/01/2010 para os mesmos horários aparecem na Figura 5.17 e Tabela 5.13, enquanto que para a tempestade do dia 26/01/2010 foram considerados os horários 16, 18, 21 e 23 UTC devido à falta de dados de radar nos horários das 14, 15 e 0 UTC, sendo os resultados expostos na Figura 5.18 e Tabela 5.14. Em termos do desempenho de classificação obtido pelo software EPPA para essas três tempestades, constatou-se que, em geral, os valores de POD ficaram próximos ou acima de 90% enquanto que os valores de FAR, próximos ou abaixo de 30%, exceto pelo primeiro

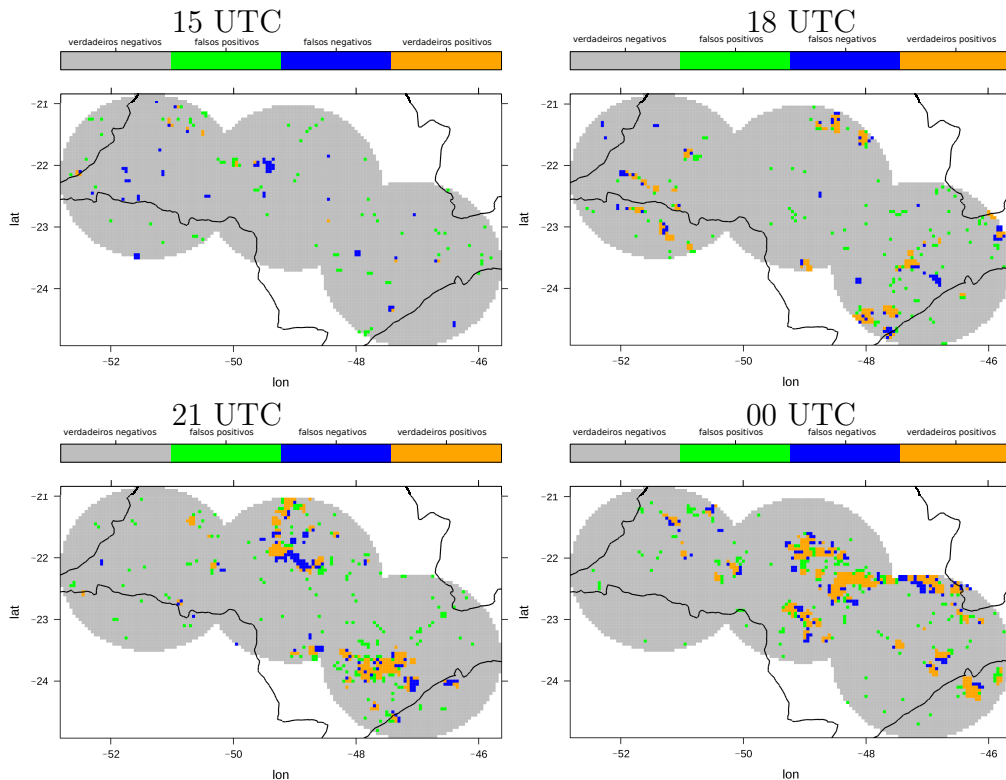


Figura 5.16 - Previsão da árvore de decisão (EPPA) relativa à ocorrência de precipitação forte ou convectiva para a tempestade do dia 14/01/2010. A sequência de imagens corresponde aos horários das 15, 18 e 21 UTC, e 0 UTC do dia seguinte.

horário da primeira tempestade.

Existe um grande número de combinações possíveis de esquemas de vizinhança, esquemas de treinamento e de ajuste de parâmetros da árvore de decisão. Os testes constituem estudos de caso da ferramenta EPPA. A implementação do software EPPA demandará a definição de um conjunto de árvores de decisão específicas para diferentes regiões ou estações do ano. Alternativamente, pode-se incluir a região e a estação do ano como atributos adicionais, o que permitiria a definição de uma única

Tabela 5.12 - Desempenho de classificação da árvore de decisão (EPPA) para previsão de ocorrência de precipitação forte e convectiva, sem vizinhança espacial ou temporal, para a tempestade do dia 14/01/2010, para as quatro previsões consideradas.

	15 UTC	18 UTC	21 UTC	00 UTC
POD	40%	88%	89%	95%
FAR	64%	27%	22%	15%

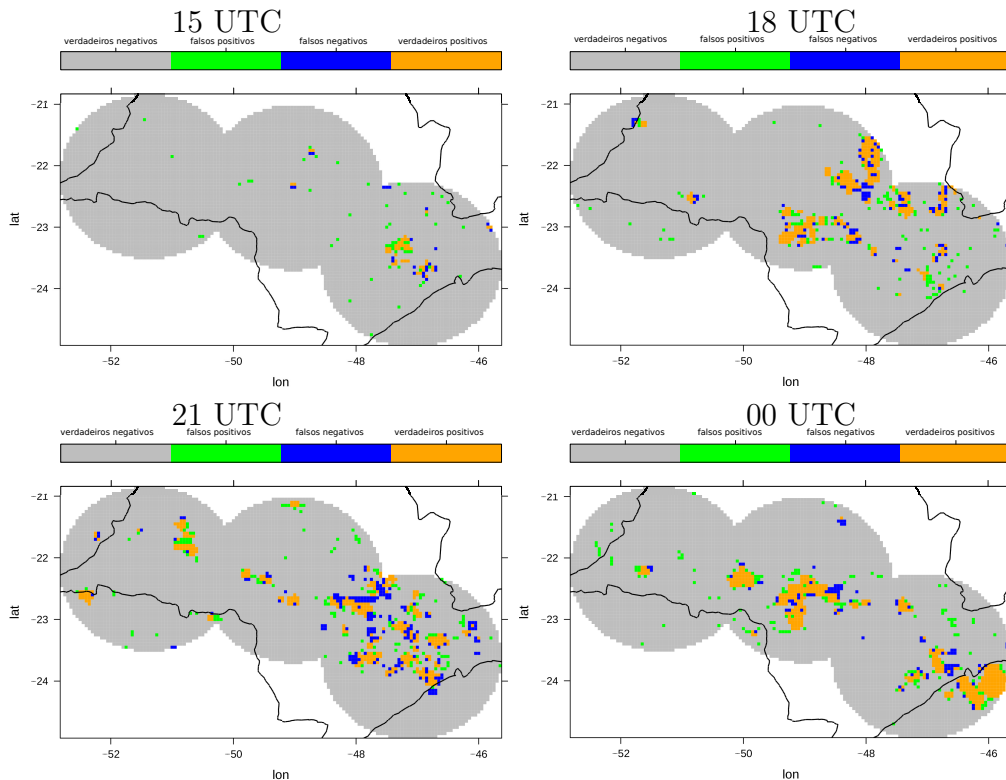


Figura 5.17 - Previsão da árvore de decisão (EPPA) relativa à ocorrência de precipitação forte ou convectiva para a tempestade do dia 19/01/2010. A sequência de imagens corresponde aos horários das 15, 18 e 21 UTC, e 0 UTC do dia seguinte.

árvore de decisão de uso geral. Conseqüentemente, torna-se necessária a realização de um conjunto de testes mais extenso, abrangendo mais dados de modelos meteorológicos. Por outro lado, uma vez que novos dados do modelo ETA 5 km só estão disponíveis para uma área bem limitada (correspondente à Serra do Mar do estado de São Paulo), será necessária a migração para dados de outros modelos, tais como o WRF.

Com relação aos possíveis esquemas de treinamento, realizou-se um teste isolado, fazendo-se o treinamento somente com os dados de análise das 00 UTC do modelo

Tabela 5.13 - Desempenho de classificação da árvore de decisão (EPPA) para previsão de ocorrência de precipitação forte e convectiva, sem vizinhança espacial ou temporal, para a tempestade do dia 19/01/2010.

	15 UTC	18 UTC	21 UTC	00 UTC
POD	93%	93%	93%	96%
FAR	31%	9%	7%	13%

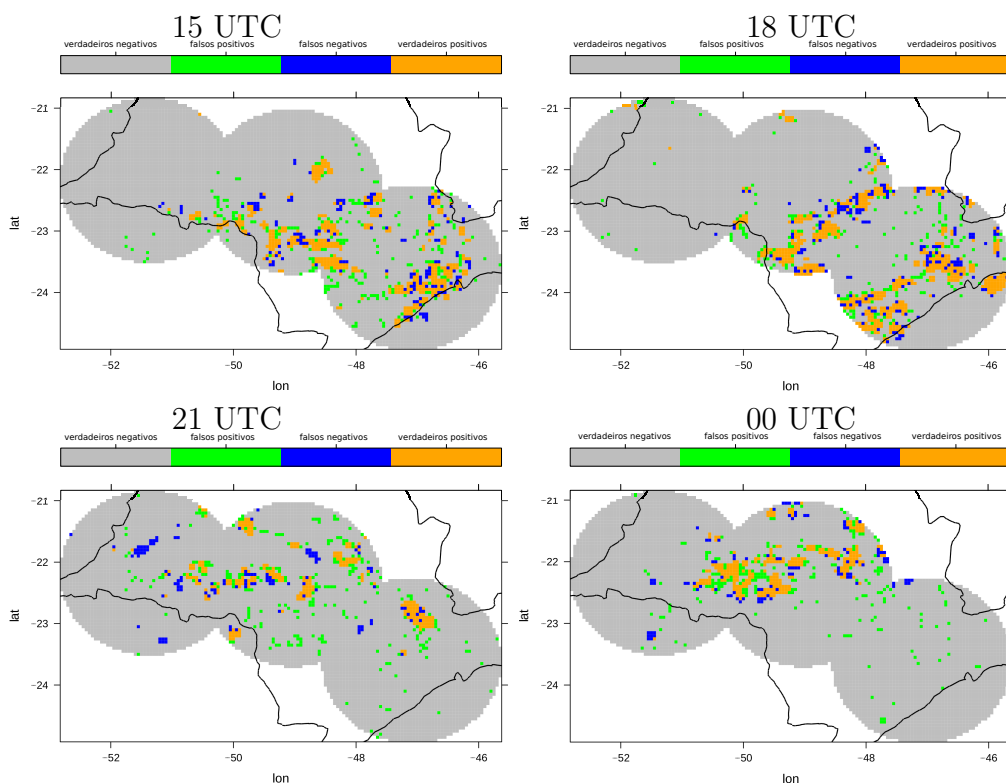


Figura 5.18 - Previsão da árvore de decisão (EPPA) relativa à ocorrência de precipitação forte ou convectiva para a tempestade do dia 26/01/2010. A sequência de imagens corresponde aos horários das 16, 18 e 21 e 23 UTC.

Tabela 5.14 - Desempenho de classificação da árvore de decisão (EPPA) para previsão de ocorrência de precipitação forte e convectiva, sem vizinhança espacial ou temporal, para a tempestade do dia 26/01/2010.

	16 UTC	18 UTC	21 UTC	23 UTC
POD	96%	96%	83%	92%
FAR	9%	7%	24%	17%

ETA 5 km, obtendo-se um bom resultado de classificação para o conjunto de treinamento: POD = 89% e FAR = 3%. Entretanto, o resultado para o conjunto de teste, que inclui previsões de 01 a 23 h, foi inaceitável: POD = 17% e FAR = 73%. Esse teste demonstra que é necessário incluir dados de previsões no treinamento da árvore de decisão.

Diversos esquemas de treinamento foram explorados, sendo expostos a seguir, sempre para dados do modelo ETA 5 km, nas Tabelas 5.15 e 5.16. Nesses testes, denomina-se *hold out* aleatório aquele em que dados de treinamento e de teste são separados aleatoriamente. O *hold out* cronológico do tipo I corresponde ao uso dos dados de

análise e de previsões até 23 h para o conjunto de treinamento, enquanto os dados das previsões de 24 a 47 h são usadas para o conjunto de testes. O *hold out* cronológico (II) corresponde a separar os dados numa proporção de 80% para treinamento e 20% para teste, mantendo a ordem cronológica dos dados. Um instante de tempo é definido para separar os dados de treinamento que o antecedem dos dados de teste, que o sucedem. Dessa forma, simula-se uma previsão. Exceto pelo *hold out* cronológico do tipo I, os dados de treinamento e teste incluem a análise das 00 UTC e previsões de 01 a 47 h.

Considerando-se os testes feitos com os dados do verão de 2009/2010, expostos na Tabela 5.15, os resultados para o *hold out* aleatório aparecem nas duas primeiras linhas. Observa-se na primeira linha que o valor de POD no teste foi relativamente baixo (69%) e que a razão entre falsos positivos e falsos negativos (FP/FN) foi muito baixa (0,35), quando estabeleceu-se que a razão mínima desejável era de 2:1, em função de ser desejável ter mais falsos positivos do que falsos negativos. Uma vez que o valor de FAR era baixo (17%), pôde-se ajustar o limiar de impureza que, como visto anteriormente, diminui os falsos negativos às custas de um aumento nos falsos positivos, e um consequente aumento do FAR. Consequentemente, conforme se observa na segunda linha da mesma tabela, o FAR aumentou para 26%, um valor razoável tendo em vista que o POD subiu para 81%.

Considerando-se ainda a Tabela 5.15, os resultados para previsão de chuva forte e convectiva usando dados do modelo ETA 5 km com *hold out* cronológico dos tipos I e II não foram satisfatórios. No caso do *hold out* cronológico do tipo II, este resultado demonstra que a árvore foi treinada parcialmente, não tendo sido treinada para todos os padrões, no caso aqueles que apareceram nos 5 dias restantes do conjunto de teste. Pode-se concluir que o conjunto de treinamento (para o *hold out* cronológico dos tipos I e II) precisaria ser expandido com mais dados de mais meses de verão, de forma a se obter uma árvore mais robusta, que consiga classificar corretamente todas as instâncias de chuva forte e convectiva. Não se tentou obter um melhor resultado variando o limiar de impureza, pois o valor de FAR já estava muito alto.

Em consequência, desses maus resultados, passou-se a utilizar dados do verão de 2010/2011 (dezembro de 2010, janeiro de 2011 e fevereiro de 2011). Além de se aumentar o volume de dados devido à utilização de mais meses, uma análise preliminar dos dados permitiu constatar que este verão teve em média 7 vezes mais tempestades que o verão de 2009/2010, constituindo assim um conjunto de dados mais representativo. Estes resultados para o verão de 2010/2011 aparecem na Ta-

bela 5.16. Observa-se nas três primeiras linhas desta tabela que o *hold out* aleatório manteve o bom desempenho. O limiar de impureza de 30% foi suficiente para atingir uma razão FP/FN de 4, enquanto que, um limiar de 10% possibilitou uma razão melhor, de 48:1. Os esquemas de *hold out* cronológico dos tipos I e II apresentaram melhora comparados aos resultados do verão de 2009/2010. O melhor resultado desta tese, correspondente ao *hold out* cronológico do tipo II, e utilizando um limiar de impureza de 5%, aparece na última linha desta tabela: POD de 84%, FAR de 55% e razão FP/FN de 8,56. O ajuste do limiar de impureza foi fundamental para um bom desempenho de classificação conforme apresentado na Tabela 5.17 que compara os índices de desempenho de classificação para vários limiares de impureza, sempre considerando-se o verão de 2010/2011 e *hold out* cronológico do tipo II.

Considerando-se o melhor resultado discutido acima, a árvore correspondente apresentou 32 níveis, correspondente ao limite máximo do pacote *rpart* do ambiente R, sendo que o número de nós resultantes foi de 512.139. Para esta mesma árvore, calculou-se a importância relativa dos atributos em termos da contribuição para a redução da impureza dos nós, sendo a soma de todas importâncias relativas correspondente a 100%. Estes valores aparecem na Tabela 5.18 para os 11 atributos considerados, observando-se que nenhum deveria ter sido desconsiderado na classificação.

Ainda para este mesmo melhor caso, a correspondente matriz de confusão é apresentada na Tabela 5.19, porém no formato que inclui o efeito da vizinhança relaxada (espacial e temporal). Nesta tabela, P e N também denotam o número de positivos e negativos. As células referentes aos falsos positivos e negativos (diagonal secundária) foram quebradas de maneira a apresentar a proporção de falsos positivos e de falsos negativos que foram reclassificados como verdadeiros positivos devido à vizinhança relaxada. Por exemplo, entre os falsos positivos (previstos e não observados) separou-se 11,1% que são vizinhos a algum pixel observado, restando 28,0% considerados como “falsos positivos relaxados”. O FAR passa a ser calculado como: $28,0 / 50,6 \times 100\% = 55\%$. Da mesma forma, entre os falsos negativos (observados e não previstos), separou-se 5,8% que são vizinhos a algum pixel previsto, restando 3,3% como “falsos negativos relaxados”. O POD passa a ser calculado como: $(11,5+5,8) / 20,6 \times 100\% = 84\%$. Esses resultados são expressos de maneira mais concisa na Tabela 5.20 que apresenta o total de verdadeiros positivos aumentado de 11,5% para 28,4%, números menores de falsos negativos e falsos positivos (3,3% e 28,0%, respectivamente) e o mesmo número de verdadeiros negativos (40,3%). Nesta tabela, observa-se que o classificador classificou corretamente 89,6% das instâncias

Tabela 5.15 - Desempenho médio da árvore de decisão (EPPA) para previsão de ocorrência de chuva forte ou convectiva, com dados do modelo ETA 5 km para o verão de 2009/2010, utilizando diversos tipos de treinamento da árvore de decisão.

<i>Hold out</i>	limiar de impureza	treinamento		teste		
		POD	FAR	POD	FAR	FP / FN
Aleatório	50%	75	7	69	17	0,35
Aleatório	10%	86	19	81	26	2,12
cronológico (I)	50%	82	5	17	68	0,81
cronológico (II)	50%	81	6	11	70	0,51

Tabela 5.16 - Desempenho médio da árvore de decisão (EPPA) para previsão de ocorrência de chuva forte ou convectiva, com dados do modelo ETA 5 km para o verão de 2010/2011, utilizando diversos tipos de treinamento da árvore de decisão.

<i>Hold out</i>	limiar de impureza	treinamento		teste		
		POD	FAR	POD	FAR	FP / FN
Aleatório	50%	94	7	93	14	1,77
Aleatório	30%	96	11	96	17	4,11
Aleatório	10%	99	27	99	30	48,22
cronológico (I)	50%	96	5	55	55	1,11
cronológico (II)	50%	96	6	54	56	1,11
cronológico (II)	5%	100	32	84	55	8,56

positivas e 59,0% das instâncias negativas, sendo o maior número de instâncias mal classificadas correspondente aos falsos positivos.

Um fator importante para o uso de árvores de decisão é a possibilidade de se obter regras ou padrões, eventualmente semelhantes à da Ferramenta Objetiva de Previsão do CPTEC/INPE, os quais são extraídos da própria estrutura da árvore resultante do treinamento. Cada sequência de condicionais, com início no nó raiz e terminando num nó terminal, dá origem a uma regra particular. Cada regras consiste num conjunto de condicionais que cada variável em particular deve atender para que a regra possa ser verificada. Nesse caso, o resultado da previsão é positivo, caso contrário (regra não atendida), o resultado é negativo. Caso haja vários condicionais para a mesma variável, adota-se aquele que for mais restritivo.

A título de exemplo, considerando-se a melhor árvore obtida, a regra mais expressiva atendeu ao maior número de instâncias positivas durante o treinamento, ou seja 5492 instâncias (0,2% do total das instâncias positivas). O nó terminal corres-

Tabela 5.17 - Desempenho médio da árvore de decisão (EPPA) para previsão de ocorrência de chuva forte ou convectiva, com dados do modelo ETA 5 km para o verão de 2010/2011, utilizando *hold out* cronológico do tipo II, para diferentes limiares de impureza.

limiar de impureza	treinamento		teste		
	POD	FAR	POD	FAR	FP / FN
50%	96	6	54	56	1,11
30%	97	10	61	56	1,61
10%	100	23	76	55	4,43
5%	100	32	84	55	8,56

Tabela 5.18 - Importância relativa (em %) de cada atributo conforme observado no treinamento da melhor árvore (*hold out* cronológico do tipo II, com limiar de impureza de 5%) com os dados do modelo ETA 5 km para o verão de 2010/2011.

TT	11	umrl ₈₀₀₋₈₅₀	9	CAPE	8
VT	10	umrl ₈₅₀₋₁₀₀₀	9	omega ₅₀₀	8
CT	10	SWEAT	9	CINE	7
K	10	BLI	9		

Tabela 5.19 - Matriz de confusão para o conjunto de teste (em %) da melhor árvore obtida (*hold out* cronológico do tipo II, com limiar de impureza de 5%) com os dados do modelo ETA 5 km para o verão de 2010/2011. As células referentes aos falsos positivos e negativos (diagonal secundária) foram quebradas de maneira a apresentar a proporção de falsos positivos e de falsos negativos que foram reclassificados como verdadeiros positivos devido à vizinhança relaxada.

		Observado			total
		P	N		
Previsto	P	11,5	11,1	28,0	Σ 50,6
	N	5,8	40,3		Σ 49,4
		3,3			
total	Σ 20,6	Σ 79,4			

Tabela 5.20 - Simplificação da tabela anterior, correspondente à melhor árvore obtida (*hold out* cronológico do tipo II, com limiar de impureza de 5%) com os dados do modelo ETA 5 km para o verão de 2010/2011.

		Observado			total
		P	N		
Previsto	P	28,4	28,0	Σ 56,4	
	N	3,3	40,3	Σ 43,6	
	total	Σ 31,7	Σ 68,3		

pondente a essa regra contém aproximadamente 25% de instâncias positivas (bem acima do limiar de impureza considerado de 5%). Esse nó terminal encontra-se no nível 28 da árvore de decisão e caracteriza um determinado padrão correspondente a instâncias consideradas como positivas. Esse padrão pode ser expresso pelos seus correspondentes condicionais, os quais podem ser interpretados como suas regras de decisão, abaixo descritas:

- $73,6 \leq \text{umrl}_{500-850} < 76,2$
- $85,6 \leq \text{umrl}_{850-1000} < 88,0$
- $20,9 \leq \text{VT} < 21,9$
- $18,4 \leq \text{CT} < 19,0$
- $\text{TT} \geq 38,6$
- $31,6 \leq \text{K} < 33,0$
- $195 \leq \text{SWEAT} < 218$
- $\text{omega}_{500} \geq -0,694$
- $73 \leq \text{CAPE} < 440$
- $\text{CINE} \geq -8$
- $-1,045 \leq \text{BLI} < -0,135$

Nas regras acima, nota-se que alguns intervalos de valores são relativamente estreitos. Isso indica que essa regra corresponde a um padrão bem específico, característico de apenas parte das instâncias positivas. Implica também que instâncias negativas que apresentavam o mesmo padrão, foram forçadamente classificadas como positivas.

A título de exemplo, foram escolhidos dois horários, ambos no dia 27 de fevereiro de 2011, sendo que no primeiro (10 UTC) observou-se grande número de pixels com chuva convectiva e forte, e no segundo (21 UTC), um grande número de pixels sem ocorrência de chuva convectiva e forte. Considerando-se que a área total dentro da cobertura dos dois radares meteorológicos abrange 4688 pixels, o primeiro horário apresentou 3014 pixels com chuva forte e convectiva, enquanto que o segundo horário, apenas 2. A melhor árvore obtida (*hold out* cronológico do tipo II, com limiar de impureza de 5%) classificou corretamente 2841 pixels dos 3014 observados como

positivos, ou seja, cerca de 94%. Para o segundo horário, esta mesma árvore classificou corretamente 3256 dos 4686 pixels observados como negativos, ou seja cerca de 69%.

Finalmente, outro aspecto importante que deve ser levado em conta é a qualidade das previsões geradas pelo modelo numérico de previsão do tempo utilizado. Essa qualidade depende da resolução do modelo e das parametrizações adotadas para o modelo, por exemplo, parametrizações relativas à transferência radiativa ou à microfísica de nuvens.

6 CONCLUSÕES E COMENTÁRIOS FINAIS

O uso de técnicas de mineração de dados em Meteorologia vem se expandindo, dada a multiplicidade e o volume crescente de dados de sensores tais como os embarcados em satélite e de modelos numéricos de previsão de tempo, executados com resoluções espaciais e temporais cada vez melhores. Neste cenário, tornam-se desejáveis ferramentas de auxílio ao meteorologista, que possam subsidiar a análise dos dados e imagens meteorológicas. Duas deficiências importantes serviram de motivação para este trabalho: (i) a pequena cobertura espacial de radares meteorológicos no Brasil e em outros países sul-americanos e africanos, e (ii) a relativa ineficiência dos modelos numéricos na previsão de precipitação convectiva.

Neste trabalho, algumas abordagens de mineração de dados para o monitoramento e previsão de atividade convectiva atmosférica a partir de dados de descargas elétricas atmosféricas foram propostas. Essas técnicas foram implementadas por dois softwares de auxílio à previsão do tempo, denominados EDDA e EPPA. O software EDDA gera campos de densidade de ocorrência de descargas elétricas NS e inclui a visualização de imagens e animações. O software EPPA objetiva realizar a previsão da ocorrência de chuva forte e convectiva a partir de previsões de um modelo numérico. Os dados meteorológicos utilizados são referentes a descargas elétricas atmosféricas, radares meteorológicos e modelos numéricos de previsão de tempo.

No escopo deste trabalho, assume-se a correlação entre descargas elétricas atmosféricas NS e atividade convectiva. Assim, o software EDDA está sendo avaliado operacionalmente no CEMADEN para o monitoramento de atividade convectiva, representando uma alternativa eficaz em relação a outros dados e produtos, na área de cobertura de sensores de detecção de descargas. Uma nova funcionalidade é também proposta para o software EDDA, a qual incorpora o agrupamento espaço-temporal das descargas NS com a geração de campos de densidade para cada grupo identificado. Essa funcionalidade permitirá identificar células eletricamente ativas de forma automático, além de fornecer parâmetros de interesse de cada célula, tais como o número de descargas, a evolução temporal da atividade elétrica, ou a carga elétrica total. Pretende-se implementar uma versão do software EDDA com agrupamento de descargas num prazo curto. Além disso, já existe uma versão do software EDDA com estimativa de precipitação convectiva a partir de dados de descargas. Essa versão está em vias de começar a ser avaliada operacionalmente no CEMADEN, sendo baseada numa função que estima a massa precipitada em função do número de descargas e fornece uma distribuição espacial da precipitação semelhante ao campo de densidade

de ocorrência de descargas NS. No entanto, esta metodologia de estimação não foi proposta ou testada no escopo desta tese.

Enquanto que o software EDDA relaciona-se mais com o monitoramento da atividade convectiva ou mesmo com *nowcasting*, o outro software proposto, o software EPPA, ainda sendo aperfeiçoado e testado, objetiva a previsão de precipitação convectiva a partir de dados de modelo numérico de previsão de tempo. As saídas do modelo sofrem uma varredura para identificação da ocorrência de padrões associados a ocorrência de atividade convectiva. Este software é baseado num algoritmo de aprendizado de máquina, a árvore de decisão e, como todo algoritmo dessa classe, inclui uma fase de treinamento e uma fase de teste. Na primeira, dados de atividade convectiva conhecidos *a priori* são utilizados para identificar os padrões, identificando-se a precipitação convectiva por meio de dados de radar. Estes padrões são constituídos de um conjunto de índices de instabilidade e variáveis atmosféricas selecionados. Nesta fase de treinamento, é gerada a árvore de decisão que será utilizada posteriormente como classificador para identificar os padrões nas saídas do modelo considerado. Adicionalmente, essa abordagem pode ser usada com qualquer modelo numérico de previsão de tempo.

Considerando-se os resultados apresentados, espera-se que os softwares EDDA e EPPA possam amenizar as duas deficiências citadas acima, ou seja, a pequena cobertura espacial dos radares meteorológicos, e a relativa ineficiência dos modelos numéricos na previsão de precipitação convectiva. Espera-se também que esses softwares representem uma contribuição para a pesquisa corrente de mineração de dados em Meteorologia.

De maneira geral, em mineração de dados, a avaliação operacional de um software aplicativo possibilita uma realimentação (*feedback*) que é muito importante para otimizar parâmetros da técnica empregada, esquemas de pre-processamento de dados, opções de visualização de resultados, ou mesmo para melhor atender aos requisitos do usuário. Assim, é natural que os softwares propostos sofram mudanças ao longo desse processo, que pode ser classificado como iterativo. No caso do software EDDA, novos estudos podem incluir uma categorização de tipos de eventos convectivos que permitiria refinar alguns parâmetros como os limiares de densidade de ocorrência ou de número de descargas.

Adicionalmente, novos dados meteorológicos podem se tornar disponíveis, constituindo uma base de dados mais robusta e que permita derivar algoritmos mais robustos, especialmente no caso do software EPPA. Observe-se que este software

utilizou dados de radar meteorológico no treinamento da árvore de decisão, mas dada a (ainda) pequena cobertura espacial destes, planeja-se, como trabalho futuro, estender esse treinamento com dados de descargas elétricas atmosféricas. Embora estes dados já estejam disponíveis para uma área muito maior que aquela da cobertura dos radares, optou-se inicialmente pelo uso dos dados de radar, devido à sua evidente maior precisão no monitoramento da atividade convectiva. Da mesma forma, no caso do software EDDA, o mapeamento da função que estima a precipitação convectiva a partir do número de descargas NS poderá ser refinado com dados combinados de precipitação estimada pelo satélite TRMM e de pluviômetros, disponíveis para acumulados diários no CPTEC/INPE. Segundo meteorologistas, estes dados são mais adequados que reanálises de modelos numéricos, em se tratando de estimação de precipitação.

Uma observação final é que uma rede de receptores de RF (rádio-frequência) para detecção de descargas representa uma opção de baixo custo em relação a uma rede extensa de radares meteorológicos para muitos países na América Latina, África e Ásia, caso se considere que softwares como os propostos possam representar uma opção para monitoramento da atividade convectiva. Obviamente essa seria uma opção de curto e médio prazo, uma vez que a longo prazo, novas gerações de satélites meteorológicos deverão suprir todo tipo de dados e imagens meteorológicos. Satélites atuais já têm, além de sensores nas bandas visíveis e infravermelhas, radares e detetores de descargas.

REFERÊNCIAS BIBLIOGRÁFICAS

- AMERICAN METEOROLOGICAL SOCIETY. **Glossary of meteorology**. 2013. Disponível em: <<http://glossary.ametsoc.or/wiki/rain>>. Acesso em: 28 de outubro 2013. 50
- ANDRADE, K. M.; MOURA, C. R. W.; ESCOBAR, G. C. J.; SILVA, P. E. D. Avaliação qualitativa do desempenho da ferramenta objetiva de previsão de tempo utilizado no ambiente operacional do CPTEC/INPE para um caso de evento severo. In: CONGRESSO BRASILEIRO DE METEOROLOGIA, 16., 2010, Belém. **Anais eletrônicos...** Belém: SBMET, 2010. Disponível em: <<http://www.cbmet2010.com/anais>>. Acesso em: 11 jul. 2012. 16, 36, 37, 49
- BARNOLAS, M.; ATENCIA, A.; LLASAT, M. C.; RIGO, T. Characterization of a mediterranean flash flood event using rain gauges, radar, GIS and lightning data. **Advances in Geosciences**, v. 17, p. 35–41, 2008. 3
- BENETI, C.; FILHO, A. J. P.; DAMIAN, E.; CALVETTI, L. Weather radar and lightning observations of mesoscale systems in the south of Brazil. In: WMO/WWRP International Symposium on Nowcasting and Very Short Range Forecasting, 3., 2012, Rio de Janeiro, Brasil. **Proceedings...** Rio de Janeiro, 2012. 4
- BETZ, H.; SCHMIDT, K.; OETTINGER, W. P.; MONTAG, B. Cell-tracking with lightning data from LINET. **Advances in Geosciences**, v. 17, p. 55–61, 2008. Disponível em: <<http://dx.doi.org/10.5194/adgeo-17-55-2008>>. Acesso em: 30 abr. 2013. 4
- BITTENCOURT, G. **Inteligência computacional**. 2013. Disponível em: <www.das.ufsc.br/~gb/pg-ic/softcomp.pdf>. Acesso em: 10 out. 2013. 29
- BLACK, T. L. The new NMC mesoscale Eta model: Description and forecast examples. **Weather Forecasting**, n. 9, p. 265–278, 1994. 16
- BONATTI, J. P. Modelo de circulação geral atmosférico do CPTEC. **Climanálise**, edição comemorativa de 10 anos, 1996. Disponível em: <<http://climanalise.cptec.inpe.br/~rclimanl/boletim/cliesp10a/bonatti.html>>. Acesso em: 1 nov. 2013. 21
- BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; STONE, C. J. **Classification and regression trees**. Belmont, CA: Wadsworth, 1984. 35

CALHEIROS, R. V.; GOMES, A. M. Flow forecasting in the Corumbataí river basin: radar rainfall stratification and runoff-rainfall relations. In: EUROPEAN CONFERENCE ON RADAR IN METEOROLOGY AND HYDROLOGY: ADVANCES IN RADAR TECHNOLOGY, 6., 2010, Sibiu, Romania.

Proceedings... [S.l.], 2010. 13, 38

CAREY, L. D.; RUTLEDGE, S. A. The relationship between precipitation and lightning in tropical island convection: a C-band polarimetric radar study.

Monthly weather review, v. 128, n. 8, p. 2687–2710, 2000. Disponível em: <[http://dx.doi.org/10.1175/1520-0493\(2000\)128<2687:TRBPAL>2.0.CO;2](http://dx.doi.org/10.1175/1520-0493(2000)128<2687:TRBPAL>2.0.CO;2)>.

Acesso em: 11 set. 2012. 2

CAVALCANTI, I. F. A. Previsão climática no CPTEC. **Climanálise**, edição comemorativa de 10 anos, 1996. Disponível em: <<http://www6.cptec.inpe.br/products/climanalise/cliesp10a/precli.html>>.

Acesso em: 1 nov. 2013. 21

_____. Large scale and synoptic features associated with extreme precipitation over South America: a review and case studies for the first decade of the 21st century. **Atmospheric Research**, v. 118, p. 27–40, 2012. 59

CECIL, D. J.; GOODMAN, S. J.; BOCCIPPIO, D. J.; ZIPSER, E. J.; NESBITT, S. W. Three years of TRMM precipitation features. Part I: radar, radiometric, and lightning characteristics. **Monthly Weather Review**, v. 133, n. 3, p. 543–566, 2005. 3

CHARBA, J. P.; SAMPLATSKY, F. G. High-resolution GFS-Based MOS quantitative precipitation forecasts on a 4-km grid. **Monthly Weather Review**, v. 139, p. 39–68, 2011. 6

CUMMINS, K. L.; MURPHY, M. J. An overview of lightning locating systems: history, techniques, and data uses, with an in-depth look at the US NLDN. **IEEE Transactions on Electromagnetic Compatibility**, v. 51, n. 3, p. 499–518, 2009. 12

DEVROYE, L. **Non-uniform random variate generation**. [S.l.]: Springer-Verlag, 1986. 27

DIXON, M.; WIENER, G. TITAN: thunderstorm identification, tracking, analysis, and nowcasting – a radar-based methodology. **Journal of Atmospheric and Oceanic Technology**, v. 10, n. 6, p. 785–797, 1993. 16

DOLIF, G.; NOBRE, C. Improving extreme precipitation forecasts in Rio de Janeiro, Brazil: are synoptic patterns efficient for distinguishing ordinary from heavy rainfall episodes? **Atmospheric Science Letters**, v. 13, p. 216–222, 2012. Disponível em: <<http://dx.doi.org/10.1002/asl.385>>. 5

EBERT, E. E.; JANOWIAK, J. E.; KIDD, C. Comparison of near-real-time precipitation estimates from satellite observations and numerical models. **Bulletin of the American Meteorological Society**, v. 88, n. 1, p. 47–64, 2007. 5

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37, 1996. ISSN 0738-4602. 30, 31

GALWAY, J. G. The lifted index as a predictor of latent instability. **Bulletin of the American Meteorological Society**, v. 37, p. 528–529, 1956. 20

GARCIA, J. V. C.; STEPHANY, S.; D'OLIVEIRA, A. B. Estimation of convective precipitation mass from lightning data using a temporal sliding-window for a series of thunderstorms in Southeastern Brazil. **Atmospheric Science Letters**, v. 14, n. 4, p. 281–286, 2013. 3, 38, 40, 41

GATLIN, P. N.; GOODMAN, S. J. A total lightning trending algorithm to identify severe thunderstorms. **Journal of Atmospheric and Oceanic Technology**, v. 27, n. 1, p. 3–22, 2010. 4

GEORGE, J. J. **Weather forecasting for aeronautics**. New York: Academic Press, 1960. 17

GLAHN, B.; GILBERT, K.; COSGROVE, R.; RUTH, D. P.; SHEETS, K. The gridding of MOS. **Weather and Forecasting**, v. 24, p. 520–529, 2009. Disponível em: <<http://dx.doi.org/10.1175/2008WAF2007080.1>>. 6

GLAHN, H. R.; LOWRY, D. A. The use of model output statistics (MOS) in objective weather forecasting. **Journal of Applied Meteorology**, v. 11, n. 8, p. 1203–1211, 1972. 6

GOODMAN, S. J. **Predicting thunderstorm evolution using ground based lightning detection networks**. [S.l.]: NASA, 1990. 216 p. (TM-10352). 48

GURGEL, H. C.; FERREIRA, N. J.; LUIZ, A. J. B. Estudo da variabilidade do NDVI sobre o Brasil, utilizando-se a análise de agrupamentos. **Revista Brasileira de Engenharia Agrícola e Ambiental**, v. 7, n. 1, p. 85–90, 2003. 5

- HAN, J.; KAMBLER, M. **Data mining**: concepts and techniques. 3. ed. New York: Elsevier, 2011. 29, 31, 48
- HARATS, N.; ZIV, B.; YAIR, Y.; KOTRONI, V.; DAYAN, U. Lightning and rain dynamic indices as predictors for flash floods events in the Mediterranean. **Advances in Geosciences**, v. 23, p. 57–64, 2010. 3
- HELD, G.; GOMES, A. M.; NACCARATO, K. P. The Structure of severe storms and associated lightning in the state of Sao Paulo, Brazil. In: EUROPEAN CONFERENCE ON RADAR IN METEOROLOGY AND HYDROLOGY: ADVANCES IN RADAR TECHNOLOGY (ERAD 2010), 6., 2010, Sibiu, Romania. **Proceedings...** [S.l.]: American Meteorological Society, 2010. 14
- HINNEBURG, A.; GABRIEL, H.-H. DENCLUE 2.0: Fast clustering based on kernel density estimation. In: BERTHOLD, M. R.; SHAW-TAYLOR, J.; LAVRA, N. (Ed.). **Advances in Intelligent Data Analysis VII**. Springer Berlin Heidelberg, 2007, (Lecture Notes in Computer Science, v. 4723). p. 70–80. ISBN 978-3-540-74824-3. Disponível em: <http://dx.doi.org/10.1007/978-3-540-74825-0_7>. ix, 32, 33, 48
- HOFFMAN, F.; JR, W. H.; III, D. E.; OGLESBY, R. Using clustered climate regimes to analyze and compare predictions from fully coupled general circulation models. **Earth Interactions**, v. 9, p. 1–27, 2005. 5
- KHAN, S.; KUHN, G.; GANGULY, A. R.; ERICKSON, D. J.; OSTROUCHOV, G. Spatio-temporal variability of daily and weekly precipitation extremes in South America. **Water Resources Research**, v. 43, n. 11, 2007. ISSN 1944-7973. Disponível em: <<http://dx.doi.org/10.1029/2006WR005384>>. 5
- KOHN, M.; GALANTI, E.; PRICE, C.; LAGOUVARDOS, K.; KOTRONI, V. Nowcasting thunderstorms in the mediterranean region using lightning data. **Atmospheric Research**, v. 100, n. 4, p. 489–502, 2011. 4
- LAKSHMANAN, V.; HONDL, K.; MACGORMAN, D.; STUMPF, G. J. The use of lightning mapping array data in WDSS-II. In: CONFERENCE ON SEVERE LOCAL STORMS, 22., 2004, Hyannis, MA. **Proceedings...** [S.l.]: American Meteorological Society, 2004. 4
- LAKSHMANAN, V.; SMITH, T. An objective method of evaluating and devising storm-tracking algorithms. **Weather and Forecasting**, v. 25, n. 2, p. 701–709, 2010. 15

LANG, T. J.; RUTLEDGE, S. A. A framework for the statistical analysis of large radar and lightning datasets: results from STEPS 2000. **Monthly Weather Review**, v. 139, n. 8, p. 2536–2551, 2011. 3, 10

LIGUORI, S.; RICO-RAMIREZ, M. A.; SCHELLART, A. N. A.; SAUL, A. J. Using probabilistic radar rainfall nowcasts and NWP forecasts for flow prediction in urban catchments. **Atmospheric Research**, v. 103, p. 80–95, 2012. 5

LIMA, G. R. T.; STEPHANY, S. A new classification approach for detecting severe weather patterns. **Computers & Geosciences**, v. 57, p. 158–165, 2013. 5, 45

_____. Training a neural network to detect patterns associated with severe weather. *Learning and Nonlinear Models (no prelo)*. 2013. 6, 45

MACHADO, L. A. T.; LIMA, W. F. A.; PINTO, O.; MORALES, C. A. Relationship between cloud-to-ground discharge and penetrative clouds: a multi-channel satellite application. **Atmospheric Research**, v. 93, n. 1, p. 304–309, 2009. 2

MATTOS, E. V.; MACHADO, L. A. T. Cloud-to-ground lightning and mesoscale convective systems. **Atmospheric Research**, v. 99, p. 377–390, 2011. 4

MEISCHNER, P. **Weather radar: principle and advanced applications**. Berlin: Springer, 2003. ISBN 3-540-000328-2. 14

MESINGER, F.; JANJIC, Z. I.; NICKOVI, S.; GAVRILOV, D.; DEAVEN, D. G. The step-mountain coordinate: model description and performance for cases of Alpine lee cyclogenesis and for a case of Appalachian redevelopment. **Monthly Weather Review**, v. 116, n. 7, p. 1493–1518, 1988. 16

MICHAELIDES, S.; SAVVIDOU, K.; NICOLAIDES, K. Relationships between lightning and rainfall intensities during rainy events in Cyprus. **Advances in Geosciences**, v. 23, p. 87–92, 2010. 3

MILLER, G. S. P. The definition and rendering of terrain maps. In: ANNUAL CONFERENCE ON COMPUTER GRAPHICS AND INTERACTIVE TECHNIQUES, 13., 1986, New York, USA. **Proceedings...** [S.l.]: ACM, 1986. p. 39–48. 26

MILLER, R. C. **Notes on analysis and severe-storm forecasting procedures of the Air Force Global Weather Central**. IL: [s.n.], 1972. 181 p. AFGWC Tech. Rep. 200 (Rev.), Air Wea. Serv., Scott AFB. 17, 18

MOSIER, R. M.; SCHUMACHER, C.; ORVILLE, R. E.; CAREY, L. D. Radar nowcasting of cloud-to-ground lightning over Houston, Texas. **Weather Forecasting**, v. 26, p. 199–212, 2011. 3, 15

NACCARATO, K. P.; PINTO, O. Improvements in the detection efficiency model for the Brazilian lightning detection network (BrasilDAT). **Atmospheric Research**, v. 91, n. 2, p. 546–563, 2009. 11

NURMI, P. **Recommendations on the verification of local weather forecasts**. [s.n.], 2003. Memorando técnico 430, 19 p. Disponível em: <<http://www.ecmwf.int/publications/library/do/references/list/14>>. Acesso em: 6 out. 2013. 50

OLIVEIRA, R. A. J.; MATTOS, E. V. The spatial-temporal relationship between cloud-to-ground lightning and precipitation distributions in the state of São Paulo. In: INTERNATIONAL CONFERENCE ON ATMOSPHERIC ELECTRICITY, 14., 2011, Rio de Janeiro, Brasil. **Proceedings...** Osaka: ICAE, 2011. 4

PESSOA, A. S. A.; LIMA, G. R. T.; SILVA, J. D. S.; STEPHANY, S.; STRAUSS, C.; CAETANO, M.; FERREIRA, N. J. Mineração de dados meteorológicos para previsão de eventos severos. **Revista Brasileira de Meteorologia**, v. 27, p. 61–74, 2012. 5, 45

PETERSEN, W. A.; RUTLEDGE, A., S.; ORVILLE, R. E. Cloud-to-ground lightning observations from TOGA COARE: selected results and lightning location algorithms. **Monthly Weather Review**, v. 124, p. 602–620, 1996. Disponível em: <[http://dx.doi.org/10.1175/1520-0493\(1996\)124<0602:CTGLOF>2.0.CO;2](http://dx.doi.org/10.1175/1520-0493(1996)124<0602:CTGLOF>2.0.CO;2)>. 2

PETERSEN, W. A.; RUTLEDGE, S. A. Regional variability in tropical convection: Observations from TRMM. **Journal of Climate**, v. 14, n. 17, p. 3566–3586, 2001. Disponível em: <[http://dx.doi.org/10.1175/1520-0442\(2001\)014<3566:RVITC0>2.0.CO;2](http://dx.doi.org/10.1175/1520-0442(2001)014<3566:RVITC0>2.0.CO;2)>. 3

PINEDA, N.; BECH, J.; RIGO, T.; MONTANYÀ, J. A Mediterranean nocturnal heavy rainfall and tornadic event. Part II: Total lightning analysis. **Atmospheric Research**, v. 100, p. 638–658, 2011. 4

PINTO, O.; PINTO, I. R. A. **Relâmpagos**. 2. ed. Sao Paulo, SP: Brasiliense, 2008. ISBN 978-85-11-00112-9. 9, 10

POLITI, J.; STEPHANY, S.; DOMINGUES, M. O.; JUNIOR, O. M. Mineração de dados meteorológicos associados a atividade convectiva empregando dados de descargas elétricas atmosféricas. **Revista Brasileira de Meteorologia**, v. 21, n. 2, p. 232–244, 2006. 25, 43

PRAKKI, S.; NOBRE, C. A.; DIAS, P. L. S. Tropics - south america. In: KAROLY, D. J.; VINCENT, D. G. (Ed.). **Meteorology of the Southern Hemisphere**. Boston: American Meteorological Society, 1998. p. 119–139. 59

R CORE TEAM. **R: a language and environment for statistical computing**. Vienna, Austria, 2013. Disponível em: <<http://www.R-project.org/>>. 50

REDE INTEGRADA NACIONAL DE DETECÇÃO DE DESCARGAS ATMOSFÉRICAS (RINDAT). **Localização dos sensores da rede RINDAT**. São José dos Campos, 2013. 1 Mapa. Disponível em: <http://www.inpe.br/webelat/rindat/imagens/Rede_RINDAT_24ss_2008.jpg>. Acesso em: 13 dez. 2013. 11

RICKENBACH, T. M.; NIETO-FERREIRA, R.; BARNHILL, R. P.; NESBITT, S. W. Regional contrast of mesoscale convective system structure prior to and during monsoon onset across South America. **Journal of Climate**, v. 24, n. 14, p. 3753–3763, 2011. 59

SCHULTZ, C. J.; PETERSEN, W. A.; CAREY, L. D. Lightning and severe weather: a comparison between total and cloud-to-ground lightning trends. **Weather and Forecasting**, v. 26, n. 5, p. 744–755, 2011. Disponível em: <<http://dx.doi.org/10.1175/WAF-D-10-05026.1>>. 4

SCOTT, D. W. **Multivariate density estimation - theory, practice and visualization**. New York: John Wiley & sons, Inc., 1992. 25

SIINGH, D.; KUMAR, P. R.; KULKARNI, M. N.; SINGH, R. P.; SINGH, A. K. Lightning, convective rain and solar activity-over the South/Southeast Asia. **Atmospheric Research**, v. 120–121, p. 99–111, 2013. 3

SILVERMAN, B. W. **Density estimation for statistics and data analysis**. London: Chapman and Hall, 1986. 23, 24, 25

STEINER, M.; JR, R. A. H.; YUTER, S. E. Climatological characterization of three-dimensional storm structure from operational radar and rain gauge data. **Journal of Applied Meteorology**, v. 34, n. 9, p. 1978–2007, 1995. 15, 42, 50

STENNING, N. V. A data transfer protocol. **Computer Networks** (1976), v. 1, n. 2, p. 99–110, 1976. 46

STRAUSS, C.; ROSA, M. B.; STEPHANY, S. Spatio-temporal clustering and density estimation of lightning data for the tracking of convective events. **Atmospheric Research**, v. 134, p. 87–99, 2013. ISSN 0169-8095. Disponível em: <<http://dx.doi.org/10.1016/j.atmosres.2013.07.008>>. 45, 58

STRAUSS, C.; STEPHANY, S. Sliding window-based spatio-temporal clustering of lightning data. In: INTERNATIONAL CONFERENCE ON ATMOSPHERIC ELECTRICITY, 14., 2011, Rio de Janeiro, Brasil. **Proceedings...** Rio de Janeiro: ICAE, 2011. 45

STRAUSS, C.; STEPHANY, S.; CAETANO, M. A ferramenta EDDA de geração de campos de densidade de descargas atmosféricas para mineração de dados meteorológicos. In: CONGRESSO NACIONAL DE MATEMÁTICA APLICADA E COMPUTACIONAL, 33., 2010, Águas de Lindóia, SP. **Anais...** São Carlos: SBMAC, 2010. v. 3, p. 269–275. ISBN 978-85-8215-040-5. 43

STRAUSS, C.; STEPHANY, S.; ROSA, M. B.; FERREIRA, N. J. Análise quantitativa das regras da ferramenta objetiva de previsão de tempo do CPTEC. In: CONGRESSO BRASILEIRO DE METEOROLOGIA, 17., 2012, Gramado, Brasil. **Anais...** SBMET, 2012. ISBN 978-85-63273-15-4. Disponível em: <<http://www.cbmet2012.com/anais/>>. 5, 50

TAPIA, A.; SMITH, J. A.; DIXON, M. Estimation of convective rainfall from lightning observations. **Journal of Applied Meteorology**, v. 37, n. 11, p. 1497–1509, 1998. 3, 40

THERNEAU, T.; ATKINSON, B.; RIPLEY, B. **rpart: recursive partitioning**. [s.n.], 2013. R package version 4.1-1. Disponível em: <<http://CRAN.R-project.org/package=rpart>>. 50

VILA, D. A.; MACHADO, L. A. T.; LAURENT, H.; VELASCO, I. Forecast and Tracking the Evolution of Cloud Clusters (ForTraCC) using satellite infrared imagery: methodology and validation. **Weather and Forecasting**, v. 23, n. 2, p. 233–245, 2008. Disponível em: <<http://journals.ametsoc.org/doi/abs/10.1175/2007WAF2006121.1>>. 5

WILLIAMS, E.; BOLDI, B.; MATLIN, A.; WEBER, M.; HODANISH, S.; SHARP, D.; GOODMAN, S.; RAGHAVAN, R.; BUECHLER, D. The behavior of

total lightning activity in severe Florida thunderstorms. **Atmospheric Research**, v. 51, n. 3, p. 245–265, 1999. 4

WITTEN, I. H.; FRANK, E. **Data mining**: practical machine learning tools and techniques. 2. ed. [S.l.]: Morgan Kaufmann Publishers, 2000. 30

ANEXO A - ARTIGOS PUBLICADOS RELACIONADOS À TESE

STRAUSS, C.; ROSA, M. B.; STEPHANY, S. Spatio-temporal clustering and density estimation of lightning data for the tracking of convective events. **Atmospheric Research**, v. 134, p. 87–99, 2013. ISSN 0169-8095. Disponível em: <<http://dx.doi.org/10.1016/j.atmosres.2013.07.008>>.

STRAUSS, C.; STEPHANY, S.; ROSA, M. B.; FERREIRA, N. J. Análise quantitativa das regras da ferramenta objetiva de previsão de tempo do CPTEC. In: CONGRESSO BRASILEIRO DE METEOROLOGIA, 17., 2012, Gramado, Brasil. **Anais...** SBMET, 2012. ISBN 978-85-63273-15-4. Disponível em: <<http://www.cbmet2012.com/anais/>>.

STRAUSS, C.; STEPHANY, S. Sliding window-based spatio-temporal clustering of lightning data. In: INTERNATIONAL CONFERENCE ON ATMOSPHERIC ELECTRICITY, 14., 2011, Rio de Janeiro, Brasil. **Proceedings...** Rio de Janeiro: ICAE, 2011.

STRAUSS, C.; STEPHANY, S.; CAETANO, M. A ferramenta EDDA de geração de campos de densidade de descargas atmosféricas para mineração de dados meteorológicos. In: CONGRESSO NACIONAL DE MATEMÁTICA APLICADA E COMPUTACIONAL, 33., 2010, Águas de Lindóia, SP. **Anais...** São Carlos: SBMAC, 2010. v. 3, p. 269–275. ISBN 978-85-8215-040-5.

Contents lists available at [SciVerse ScienceDirect](#)

Atmospheric Research

journal homepage: www.elsevier.com/locate/atmos

Spatio-temporal clustering and density estimation of lightning data for the tracking of convective events



Cesar Strauss*, Marcelo Barbio Rosa, Stephan Stephany

National Institute for Space Research (INPE), Av. dos Astronautas, 1758, Sao Jose dos Campos, SP, Brazil

article info

Article history:

Received 10 January 2013

Received in revised form 25 May 2013

Accepted 10 July 2013

Available online 18 July 2013

Keywords:

Tracking

Lightning

Convective thunderstorm

Clustering

Kernel estimation

abstract

Convective cells are cloud formations whose growth, maturation and dissipation are of great interest among meteorologists since they are associated with severe storms with large precipitation structures. Some works suggest a strong correlation between lightning occurrence and convective cells. The current work proposes a new approach to analyze the correlation between precipitation and lightning, and to identify electrically active cells. Such cells may be employed for tracking convective events in the absence of weather radar coverage. This approach employs a new spatio-temporal clustering technique based on a temporal sliding-window and a standard kernel density estimation to process lightning data. Clustering allows the identification of the cells from lightning data and density estimation bounds the contours of the cells. The proposed approach was evaluated for two convective events in Southeast Brazil. Image segmentation of radar data was performed to identify convective precipitation structures using the Steiner criteria. These structures were then compared and correlated to the electrically active cells in particular instants of time for both events. It was observed that most precipitation structures have associated cells, by comparing the ground tracks of their centroids. In addition, for one particular cell of each event, its temporal evolution was compared to that of the associated precipitation structure. Results show that the proposed approach may improve the use of lightning data for tracking convective events in countries that lack weather radar coverage.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

Convective cells are cloud formations whose growth, maturation and dissipation are associated with large precipitation structures. The current work proposes a new approach that combines a spatio-temporal method and kernel density estimation to identify, monitor and track electrically active cells of cloud-to-ground (CG) lightning occurrences. This approach is based on the assumption that CG lightning can be correlated to convective activity.

The spatio-temporal clustering of lightning data proposed here is based on a new temporal sliding-window approach (Strauss and Stephany, 2011), while the standard kernel density estimation is a well known technique (Silverman,

1986; Scott, 1992) frequently employed for lightning data. See for instance Steinacker et al. (2000) or Tuomi and Larjavaara (2005). In particular, the EDDA kernel estimation software (Strauss et al., 2010; Pessoa et al., 2012), that stands for Atmospheric Discharge Density Estimator, in Portuguese, is currently being evaluated for operational use in tracking of convective precipitation (without using clustering) in a recently established Center for Natural Disaster Monitoring and Alert (CEMADEN) in Brazil. Its usefulness is expected to improve with the addition of clustering. Kernel density estimation was also employed with a numerical weather model to develop supervised classification methods in an attempt to perform the prediction of severe convective events (Pessoa et al., 2012; Lima and Stephany, 2013).

Clustering allows the identification of the strokes that compose each electrically active cell from lightning data while density estimation bounds the corresponding borders. The proposed approach was evaluated for two convective events in

* Corresponding author. Tel./fax: + 55 12 3208 7141.

E-mail addresses: cstrauss@cea.inpe.br (C. Strauss), marcelo.barbio@cptec.inpe.br (M.B. Rosa), stephan@ac.inpe.br (S. Stephany).

ANÁLISE QUANTITATIVA DAS REGRAS DA FERRAMENTA OBJETIVA DE PREVISÃO DE TEMPO DO CPTEC

Cesar STRAUSS^{1,2}, Stephan STEPHANY³, Marcelo B. ROSA⁴, Nelson J. FERREIRA⁴

¹CEA/INPE – São José dos Campos – São Paulo - cstrauss@cea.inpe.br

³LAC/INPE – São José dos Campos – São Paulo

⁴ CPTEC/INPE – Cachoeira Paulista– São Paulo

RESUMO: O Grupo de Previsão do Tempo do CPTEC/INPE utiliza como auxílio na previsão uma ferramenta objetiva que gera cartas para visualizar variáveis selecionadas de previsões do modelo numérico regional ETA. Essas cartas visam identificar a possibilidade de ocorrência de eventos tais como pancada de chuva com trovoadas, tempestade e granizo. A seleção das variáveis e os limiares adotados para esses 3 tipos de eventos foi realizada com base na experiência dos meteorologistas da previsão de tempo e em valores de referência citados na literatura da área. O presente trabalho apresenta uma análise quantitativa do grau de acerto das variáveis e limiares selecionados utilizando dados de radares meteorológicos como referência em termos de precipitação.

ABSTRACT: The Weather Forecast Group of CPTEC/INPE employs as an ancillary scheme given by an objective forecast tool. It generates charts to visualize selected variables of forecasts of the ETA regional numerical model. These charts are intended to identify the possibility of occurrence of events such as rainshowers with thunderstorms, storms and hailstorms. The selection of the variables and the adopted thresholds for these 3 types of events were done based on the experience of weather forecast meteorologists and on values cited in literature references. The current work presents a quantitative analysis of the hit ratio of the selected variables and thresholds employing weather radar data as reference for the precipitation.

1 – INTRODUÇÃO

O Grupo de Previsão do Tempo do CPTEC/INPE idealizou uma ferramenta objetiva de auxílio na previsão chamada ferramenta objetiva de previsão que gera cartas para visualizar variáveis selecionadas de previsões do modelo numérico regional ETA com resolução de 20 km (ANDRADE ET AL., 2010). Essas cartas visam identificar a possibilidade de ocorrência de eventos tais como pancada de chuva com trovoadas, tempestade e granizo. Em cada carta, os

Sliding window-based spatio-temporal clustering of lightning data

Cesar Strauss.¹, Stephan Stephany²

1. Associated Laboratory of Computing and Applied Mathematics, National Institute for Space Research,
12245-970, S.J Campos, SP - Brazil

2. Coordination of Space and Atmospheric Sciences, National Institute for Space Research,
12245-970, S. J Campos, SP - Brazil

ABSTRACT: In this work, we introduce a novel way of clustering spatio-temporal data based on a temporal sliding window. The sliding window algorithm is employed, for instance, for data flow control in networks. The spatio-temporal clustering was applied to lightning data in order to track the evolution of the electrical activity associated to convective storms in space and time. As a result, it is possible to define these clusters as being nuclei of atmospheric electrical activity and to calculate the corresponding position of the centroid, the number of lightnings, or the neutralized electrical charge.

1. INTRODUCTION

Convective nuclei are cloud formations whose birth, evolution and dissipation are of great interest among meteorologists since they are associated to severe storms. Lightning data can be a way to indirectly track these nuclei. Lima et al. [2006] found a strong correlation between the probability of occurrence of lightnings and the WV-IR index (water vapor minus infrared index) based on GOES satellite images that show cloud tops. Caetano et al. [2009] found a similar correlation comparing the same type of GOES images to the field of density of occurrences of lightning calculated by a kernel estimation technique. The more intense regions of this field correspond to the so-called “nuclei of electrical activity”. Further work [Caetano et al., 2010] suggested a further correlation with precipitation structures in weather radar images. However, the tracking of the evolution of convective activity using lightning data needs to be better analyzed. We propose a new technique for spatio-temporal clustering applied to lightning data as an alternative or a complement to the use of the field of density of occurrences aiming at a deeper investigation of the correlation between convective nuclei and lightnings.

2. PROPOSED SPATIO-TEMPORAL CLUSTERING METHOD

The proposed method is based on a temporal sliding window, similarly as the sliding window employed for data flow control in networks. A fixed-width temporal window is *slided* in time with a constant rate, but for discrete timesteps, in order to process new incoming lightning data. The clusters obtained by this method are also called “nuclei of electrical activity”. A major difficulty that we have in the clustering process is how to identify and track a particular nucleus across space and time, while maintaining its identity. After clustering the 2D lightning data defined by a given timestep of the window, the current clusters are a mixture of new clusters and ones that already were defined in a previous timestep of the window.

*Correspondence to:

A Ferramenta EDDA de Geração de Campos de Densidade de Descargas Atmosféricas para Mineração de Dados Meteorológicos

Cesar Strauss

INPE - Coordenação de Ciências Espaciais e Atmosféricas (CEA)
12227-010, São José dos Campos, SP
E-mail: cstrauss@cea.inpe.br

Stephan Stephany

INPE - Laboratório Associado de Computação e Matemática Aplicada (LAC)
12227-010, São José dos Campos, SP
E-mail: stephan@lac.inpe.br

Mirian Caetano

INPE - Centro de Previsão de Tempo e Estudos Climáticos (CPTEC)
12630-000, Cachoeira Paulista, SP
E-mail: miriam.caetano@cptec.inpe.br

Resumo: A ferramenta EDDA estima a densidade de ocorrência de descargas elétricas atmosféricas a partir do registro de eventos individuais, para uma extensão geográfica e intervalo de tempo selecionados. O método utilizado é o estimador de núcleo gaussiano com janela adaptativa, sendo gerados arquivos em formato ASCII adequados a algoritmos de mineração e em formato de grade binário para a ferramentas de visualização meteorológica GRADS. Parâmetros específicos podem ser ajustados de forma a se poder correlacionar a densidade com outros dados meteorológicos. A ferramenta permite também visualizar animações varrendo os registros de descargas por meio de uma janela deslizante, de forma a acompanhar a evolução temporal de estruturas convectivas. A pronta disponibilidade de dados de descargas permite animações que permitem monitorar os eventos meteorológicos com atividade elétrica em tempo quase-real.

Palavras-chave: mineração de dados, previsão meteorológica, eventos convectivos, estimador de densidade

1 Introdução

Este trabalho apresenta a ferramenta EDDA, que estima a densidade de ocorrência de descargas elétricas atmosféricas para uma extensão geográfica e intervalo de tempo selecionados. Os dados brutos de descargas, contendo os registros individuais em formato ASCII são gerados pela Rede Integrada Nacional de Detecção de Descargas Atmosféricas (RINDAT), fornecidos pelo CPTEC/INPE. A ferramenta implementa o estimador de núcleo gaussiano com janela adaptativa, sendo gerados arquivos em formato ASCII adequados a algoritmos de mineração e em formato de grade binário para a ferramentas de visualização meteorológica GRADS. Parâmetros específicos podem ser ajustados de forma a se poder correlacionar a densidade com outros dados, objetivando seu uso na mineração de dados meteorológicos. A ferramenta permite também visualizar animações varrendo os registros de descargas por meio de uma janela deslizante, de forma a acompanhar a evolução temporal de estruturas convectivas. A pronta disponibilidade de dados de descargas permite animações que permitem monitorar os eventos meteorológicos com

PUBLICAÇÕES TÉCNICO-CIENTÍFICAS EDITADAS PELO INPE

Teses e Dissertações (TDI)

Teses e Dissertações apresentadas nos Cursos de Pós-Graduação do INPE.

Manuais Técnicos (MAN)

São publicações de caráter técnico que incluem normas, procedimentos, instruções e orientações.

Notas Técnico-Científicas (NTC)

Incluem resultados preliminares de pesquisa, descrição de equipamentos, descrição e ou documentação de programas de computador, descrição de sistemas e experimentos, apresentação de testes, dados, atlas, e documentação de projetos de engenharia.

Relatórios de Pesquisa (RPQ)

Reportam resultados ou progressos de pesquisas tanto de natureza técnica quanto científica, cujo nível seja compatível com o de uma publicação em periódico nacional ou internacional.

Propostas e Relatórios de Projetos (PRP)

São propostas de projetos técnico-científicos e relatórios de acompanhamento de projetos, atividades e convênios.

Publicações Didáticas (PUD)

Incluem apostilas, notas de aula e manuais didáticos.

Publicações Seriadas

São os seriados técnico-científicos: boletins, periódicos, anuários e anais de eventos (simpósios e congressos). Contam destas publicações o Internacional Standard Serial Number (ISSN), que é um código único e definitivo para identificação de títulos de seriados.

Programas de Computador (PDC)

São a seqüência de instruções ou códigos, expressos em uma linguagem de programação compilada ou interpretada, a ser executada por um computador para alcançar um determinado objetivo. Aceitam-se tanto programas fonte quanto os executáveis.

Pré-publicações (PRE)

Todos os artigos publicados em periódicos, anais e como capítulos de livros.