UNIVERSIDADE FEDERAL DO PARANÁ

MARIANA KLEINA

IDENTIFICAÇÃO, MONITORAMENTO E PREVISÃO DE TEMPESTADES ELÉTRICAS

CURITIBA
2015

MARIANA KLEINA

IDENTIFICAÇÃO, MONITORAMENTO E PREVISÃO DE TEMPESTADES ELÉTRICAS

Tese apresentada ao Curso de Pós-Graduação em Métodos Numéricos em Engenharia, Área de Concentração em Programação Matemática, do Departamento de Matemática, Setor de Ciências Exatas e do Departamento de Construção Civil, Setor de Tecnologia, Universidade Federal do Paraná, como parte das exigências para a obtenção do título de Doutor em Ciências.

Orientador: Prof. Dr. Luiz Carlos Matioli

CURITIBA
2015

K64i Kleina, Mariana
Identificação, monitoramento e previsão de tempestades elétricas/ Mariana Kleina. – Curitiba, 2015.
112 f. : il. color. ; 30 cm.

Tese - Universidade Federal do Paraná, Setor de Ciências Exatas, Programa de Pós-graduação em Métodos Numéricos em Engenharia, 2015.

Orientador: Luiz Carlos Matioli .
Bibliografia: p. 103-107.

1. Análise por conglomerados. 2. Raio. 3. Previsão do tempo. I. Universidade Federal do Paraná. II.Matioli, Luiz Carlos. III. Título.

CDD: 551.563019

Ministério da Educação
Universidade Federal do Paraná
Setor de Tecnologia / Setor de Ciências Exatas
Departamento de Construção Civil / Departamento de Matemática/ Departamento de Engenharia de Produção.
Programa de Pós-Graduação em Métodos Numéricos em Engenharia – PPGMNE/UFPR.

TERMO DE APROVAÇÃO

MARIANA KLEINA

IDENTIFICAÇÃO, MONITORAMENTO E PREVISÃO DE TEMPESTADES ELÉTRICAS

Tese aprovada como requisito parcial para obtenção do grau de doutora no Programa de Pós-Graduação em Métodos Numéricos em Engenharia, da Universidade Federal do Paraná, pela seguinte banca examinadora:

Prof. Dr. Luiz Carlos Matioli
Orientador – Membro do PPGMNE/UFPR

Prof. Dr. Sérgio Scheer
Membro do PPGMNE/UFPR

Dr. Eduardo Alvim Leite
Membro do SIMEPAR

Dr. Leonardo Calvetti
Membro do SIMEPAR

Prof Dr. Cleverson Luiz da Silva Pinto
Membro da COPEL

Curitiba, 10 de dezembro de 2015.

AGRADECIMENTOS

A Deus, por absolutamente tudo.

A minha família, pelo amor e apoio incondicional.

Ao Matioli e ao Alvim, pelo conhecimento transmitido.

Ao PPGMNE, UFPR e SIMEPAR, pela oportunidade que me foi dada.

Aos meus amigos, pelos momentos inesquecíveis.

A todos que contribuíram para a realização deste trabalho.

O sucesso nasce do querer, da determinação e persistência em se chegar a um objetivo. Mesmo não atingindo o alvo, quem busca e vence obstáculos, no mínimo fará coisas admiráveis.

José de Alencar

RESUMO

Este trabalho tem por objetivo propor um sistema de identificação, monitoramento e previsão de tempestades elétricas em uma região de linha de transmissão de energia do Brasil utilizando apenas informações de descargas atmosféricas. Na etapa de identificação das tempestades é utilizada clusterização espacial/temporal das descargas, para o monitoramento são realizadas conexões entre *clusters* de acordo com a velocidade de deslocamento dos núcleos das tempestades e finalmente a previsão uma hora à frente é obtida por meio de extrapolação de dados. As variáveis das tempestades elétricas acompanhadas e previstas são: posição central (latitude e longitude), número médio de descargas por hora, distribuição espacial das descargas dentro das tempestades e média do valor absoluto do pico de corrente de descargas que compõem as tempestades. O sistema é calibrado por meio da otimização de um problema irrestrito baseado em índices de monitoramento e previsão. Exemplos práticos são apresentados para a região piloto, onde diversos gráficos ilustram o produto do sistema proposto.

Palavras-chave: Análise de Agrupamento. *Convergent Data Sharpening*. Descargas Atmosféricas. Tempestades Elétricas. Previsão à Curto Prazo.

ABSTRACT

This work aims to propose a system of identification, monitoring and forecasting of electrical storms on a region of Brazilian power transmission line using lightning information only. In the step of storms identification is used spatial/temporal lightning clustering, to monitoring are performed connections between clusters according to the displacement speed of the storms cores and finally the forecast one hour ahead is obtained by data extrapolation. The monitored and forecasted variables of the electrical storms are: center position (latitude and longitude), average number of lightning per hour, spatial distribution of lightning into the storms and average of the absolute value of the peak current of lightning that comprise the storms. The system is calibrated by optimizing an unrestricted problem based on monitoring and forecasting indexes. Practical examples are presented to the pilot region and several graphs illustrate the proposed system product.

Key-words: Cluster Analysis. Convergent Data Sharpening. Lightning Strokes. Electrical Storms. Forecast Short Term.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIAÇÕES

AGNES: *Agglomerative Nesting*
ANEEL: Agência Nacional de Energia Elétrica
CEMIG: Companhia Energética de Minas Gerais
CLIQUE: *Clustering In Quest*
COPEL: Companhia Paranaense de Energia Elétrica
DBSCAN: *Density Based Spatial Clustering of Applications with Noise*
DENCLUE: *Density-based Clustering*
DIANA: *Divisive Analysis*
INPE: Instituto Nacional de Pesquisas Espaciais
KDE: *Kernel Density Estimation*
RINDAT: Rede Integrada Nacional de Detecção de Descargas Atmosféricas
SIMEPAR: Sistema Meteorológico do Paraná
TITAN: *Thunderstorm Identification, Tracking, Analysis and Nowcasting*
TRT: *Thunderstorms Radar Tracking*
SOM: *Self-Organizing Map*
STING: *Statistical Information Grid*
UTC: *Universal Time Coordinate*

SUMÁRIO

1 INTRODUÇÃO

O Brasil é líder mundial na ocorrência de descargas atmosféricas com cerca de 50 a 60 milhões de ocorrências por ano (INPE, 2015). Estas são produzidas devido ao acúmulo de cargas elétricas na atmosfera geradas por sistemas meteorológicos de diversos tipos e configurações, denominados genericamente de tempestades. Descargas atmosféricas, mais especificamente as que atingem o solo, chamadas descargas nuvem-solo, despertam interesse de estudo pois são responsáveis tanto por afetar a segurança humana, provocando mortes e injúrias, quanto por grandes prejuízos em diversos setores econômicos, com destaque para o setor elétrico. Neste setor, elas são responsáveis por danos nos equipamentos e desligamentos não programados nos sistemas de transmissão e distribuição de energia, podendo gerar apagões de grande abrangência e impactos econômicos.

Descargas atmosféricas são fenômenos naturais associadas a eventos meteorológicos que podem durar de poucas horas até vários dias. Quando organizadas, as descargas configuram tempestades elétricas, as quais podem causar perigo eminente às áreas por onde se propagam. Se monitoradas, diversos atributos das tempestades podem ser estimados e utilizados para categorizar um evento atuante.

Tempestades elétricas, nesta pesquisa, serão caracterizadas por meio da clusterização de descargas atmosféricas considerando as variáveis espaço e tempo. Por intermédio da clusterização de dados, é possível revelar e interpretar grande quantidade de informações por meio de atributos comuns que dados do mesmo grupo possuem. Assim, “a clusterização é um procedimento exploratório que busca por uma estrutura natural com relação ao conjunto específico, que envolve ordenar os dados em grupos tal que os objetos no mesmo *cluster* são mais parecidos entre si do que em relação aos objetos em outro *cluster*” (CARLANTONIO, 2001). Desta forma, nesta pesquisa,

uma tempestade elétrica representa um conjunto de descargas próximas espacialmente e temporalmente. O processo de repetição da clusterização das descargas em intervalos de tempos regulares juntamente com conexões apropriadas entre *clusters* permite acompanhar o surgimento, a maturação e a dissipação, bem como calcular atributos ao longo da vida de tempestades elétricas.

Um bom monitoramento é parte primordial para outra etapa de suma importância na vigilância de tempestades elétricas: a previsão de seus atributos. Prever algo geralmente não é uma tarefa fácil, especialmente quando estão envolvidos fenômenos naturais e muitas incertezas associadas ao processo. Para previsões confiáveis, diversos requisitos devem ser cumpridos, entre eles, o conhecimento amplo do problema, coleta de dados acurados, escolha de um bom modelo de previsão que atenda as expectativas, entre outros. A confiança em um prognóstico pode auxiliar em tomadas de decisões importantes, especialmente quando há riscos envolvidos.

A previsão de atributos das tempestades elétricas, nesta pesquisa, será realizada por extrapolação de dados, que basicamente consiste em aproximar um valor desconhecido fora de um intervalo de pontos conhecidos, visto que primeiramente uma curva é ajustada aos pontos dados. Extrapolação de dados é uma técnica bastante utilizada para realizar previsão de variáveis meteorológicas, podendo-se citar os trabalhos de Dixon e Wiener (1993), Hering *et al.* (2004) e Bonelli e Marcacci (2008). Tendo o conhecimento prévio do comportamento de uma tempestade eletricamente ativa que se desloca para uma região de risco, como é o caso de áreas com linhas de energia, medidas operacionais podem ser executadas, reduzindo, por exemplo, custos com consertos de equipamentos e multas por falta de energia para as concessionárias.

1.1 ESTUDOS PRELIMINARES

Antes da idealização do sistema de detecção, monitoramento e previsão de tempestades elétricas apresentado nesta pesquisa, clusterização de descargas atmosféricas foi estudado por Kleina, Matioli e Leite (2014) que analisaram a variabilidade da

intensidade do pico de corrente de descargas atmosféricas a nível de cada evento de tempestade, formadas por técnicas de clusterização. Buscou-se identificar se o comportamento do pico de corrente apresenta diferença em relação ao comportamento climatológico, o qual se considera a distribuição geral de probabilidade, estimado a partir de medições em todo o planeta. Para isso, foram tomadas como base as descargas elétricas incidentes no ano de 2011 na região da linha de transmissão de energia LT 765 kV (a mesma região do presente estudo, apresentada na Seção 2.4.1). Estas descargas foram então agrupadas por diversas técnicas de clusterização de dados e gerados índices estatísticos próprios. Os resultados obtidos comprovaram o pressuposto de que tempestades elétricas se caracterizam como tipo específico de eventos meteorológicos, estão sob influência de condições atmosféricas próprias, assim apresentando comportamentos específicos de descargas em relação ao seu pico de corrente, diferindo do comportamento climatológico conhecido na literatura.

Em Kleina *et al.* (2015), tempestades elétricas foram identificadas por meio de clusterização de descargas no espaço e no tempo, buscando analisar variáveis meteorológicas destas tempestades e associá-las à perturbações na LT 765 kV. A clusterização foi realizada pelo método *Convergent Data Sharpening*. Três cenários de estudo foram elaborados para análise: o primeiro levou em conta todas as tempestades identificadas na região de estudo, o segundo englobou tempestades relativamente próximas à linha de energia e o terceiro incluiu tempestades que continham descargas associadas à falhas no sistema elétrico. Para cada cenário, foram gerados padrões de tempestades pelo método *Self-Organizing Map* (SOM) a fim de representar o grande volume de tempestades elétricas identificadas em 14 anos de dados. Na criação dos preditores para entrada no método SOM, um polinômio foi ajustado para cada trajetória de tempestade, assim cada padrão gerado pelo SOM representou um conjunto de tempestades com trajetórias similares. O pico de corrente médio e o número de descargas dos padrões de tempestades foram investigadas com o intuito de averiguar se estas possuem relação com falhas no sistema elétrico considerando os cenários criados. Com relação a variável pico de corrente, notou-se que tempestades em situações

de falhas tem magnitude similar a situações onde não se observaram falhas, sendo sudoeste da região que engloba a linha de energia analisada a área com maior intensidade desta variável. Porém, quando há falhas no sistema elétrico, as tempestades apresentam maior número de descargas do que em situações sem falhas.

Estes dois estudos citados anteriormente foram primordiais para concepção da presente pesquisa. Eles serviram de grande aprendizado e familiaridade com os métodos numéricos que serão utilizados, especialmente o algoritmo de clusterização, e também em relação aos dados de aplicação os quais são as descargas atmosféricas e consequentemente as tempestades elétricas. O estudo que será apresentado nesta pesquisa pode ser encontrado em Kleina, Matioli e Leite (2015).

1.2 MOTIVAÇÃO

Devido aos números relevantes de prejuízos causados por descargas atmosféricas em diversos setores econômicos, em especial ao setor elétrico, cada vez mais busca-se conhecer e entender melhor esse fenômeno que pode causar enormes danos, não somente materiais.

O setor elétrico é bastante impactado por descargas, especialmente no Brasil, país líder na incidência de tais fenômenos, devido ao seu clima tropical e sua grande extensão territorial. Minimizar, ao menos um pouco, estes impactos pode representar melhorias significativas, tanto para consumidores quanto para as empresas de energia elétrica, reduzindo os indesejáveis desligamentos não programados quando causados por esse tipo de fenômeno.

A elaboração de um sistema que monitore e realize previsões de variáveis meteorológicas (como o pico de corrente de descargas de tempestades, que apresenta relação significativa com falhas e interrupções) pode representar uma ferramenta diferenciável que auxilie especialistas quando tomada de decisões são necessárias.

1.3 OBJETIVOS

1.3.1 Objetivo Geral

O objetivo principal desta pesquisa é o desenvolvimento de um sistema de detecção, monitoramento e previsão a curto prazo de tempestades elétricas, baseado apenas em dados de descargas nuvem-solo, utilizando métodos numéricos. Tal sistema permitirá acompanhar a evolução de uma tempestade elétrica que se desenvolve na região piloto, em um certo período, além da previsão de uma hora à frente de seus atributos.

Dispor do conhecimento de trajetórias e atributos de tempestades elétricas, tanto do passado quanto do futuro, pode significar um grande avanço para o desenvolvimento de sistemas de vigilância e alertas, em tempo real, em regiões que envolvem algum tipo de risco, como é o caso de áreas que abrangem linhas de energia.

1.3.2 Objetivos Específicos

- Identificar núcleos de tempestades elétricas utilizando clusterização espacial e temporal de descargas atmosféricas;
- Conectar núcleos de tempestades (*clusters*) de períodos consecutivos por meio da velocidade de deslocamento, a fim de representar o movimento das tempestades elétricas;
- Realizar previsões uma hora à frente dos atributos das tempestades elétricas e analisar suas peculiaridades, especialmente quando rumam em direção à linha de energia LT 765 kV.

1.4 INOVAÇÕES PROPOSTAS

Esta pesquisa tem por inovação a caracterização de tempestades elétricas utilizando clusterização de descargas atmosféricas para a formação de núcleos de ativi-

dade elétrica, e encadeamento desses núcleos no tempo para representar o seguimento das tempestades. Estes dados contemplam apenas a posição espacial e temporal das descargas, que quando organizadas caracterizam tempestades elétricas. Geralmente imagens de radar são utilizadas para identificar, monitorar e prever células de tempestades, e eventualmente dados de descargas atmosféricas complementam informações e/ou são utilizados para preencher lacunas quando ocorrem falhas com os demais dados, já que áreas de alta refletividade (e consequentemente precipitação) são geralmente áreas de alta densidade de descargas (BONELLI; MARCACCI, 2008; SOULA, 2009).

A clusterização de dados, nesta pesquisa, procura estruturar espacial e temporalmente as descargas a fim de caracterizar o fenômeno meteorológico que as governam, assim descargas que são próximas no espaço (latitude/longitude) e no tempo (uma hora) são consideradas pertencentes ao mesmo evento. A clusterização é realizada pelo método *Convergent Data Sharpening*, que é uma técnica bastante propícia para este problema pois identifica como núcleo de uma tempestade elétrica o valor modal dos dados de um *cluster*, e a conectividade com o(s) núcleo(s) do período seguinte é muito mais realista pois acompanha o movimento majoritário das descargas. Após o monitoramento, diversos atributos das tempestades podem ser calculados e projetados a curto prazo.

O sistema proposto utiliza apenas informações de descargas atmosféricas (não faz uso de outras informações meteorológicas provindos de radares e satélites, por exemplo) e não necessita de um grande volume de dados. O sistema precisa apenas de um pequeno período de treinamento para organizar as descargas e monitorar as tempestades, para então prever atributos das tempestades.

Outro diferencial do sistema proposto é sua capacidade de fazer previsão de uma variável muito importante e bastante conhecida pelo seu potencial de impacto da descarga quando atinge o solo: o pico de corrente.

1.5 ORGANIZAÇÃO DO TRABALHO

No Capítulo 2 são abordadas definições e conceitos das descargas atmosféricas. No Capítulo 3 é apresentado o embasamento teórico de agrupamentos de dados e suas diversas categorias e também é descrito o algoritmo de clusterização utilizado na pesquisa. No Capítulo 4 é apresentada a forma de aplicação para o desenvolvimento do sistema de detecção, monitoramento e previsão de tempestades elétricas, assim como a escolha do parâmetro do sistema. Já no Capítulo 5, o sistema proposto é apresentado de forma pré-operacional, onde são apresentados alguns casos particulares para exemplificar os resultados do sistema. As conclusões e sugestões para trabalhos futuros são apresentadas no Capítulo 6. Finalmente, no Apêndice A constam mais exemplos visuais dos resultados do sistema proposto na pesquisa.

2 DESCARGAS ATMOSFÉRICAS

Neste capítulo serão apresentados conceitos básicos sobre descargas atmosféricas: sua formação, características e diversas outras definições associadas a elas. Também são abordados os impactos causados por descargas, especialmente ao setor elétrico, em que um caso particular deste é foco da aplicação da pesquisa.

2.1 CONCEITOS E DEFINIÇÕES

Nesta seção será apresentado conceito do fenômeno descarga atmosférica, conforme as referências: Golde (1997), Macgorman e Rust (1998), Uman (2001), Rakov e Uman (2003), Heidler *et al.* (2008) e Instituto Nacional de Pesquisas Espaciais (INPE, 2015).

2.1.1 O que são as Descargas Atmosféricas

Descargas atmosféricas são descargas elétricas de grande intensidade (pico de corrente – valor máximo atingido pela corrente – acima de um quiloàmpere (kA)) e de grande extensão (alguns quilômetros). Inicia-se quando é excedida a capacidade isolante (ou rigidez dielétrica) do ar pelo campo elétrico gerado pelo acúmulo de cargas elétricas em regiões localizadas na atmosfera. Quebrada a rigidez dielétrica do ar, elétrons de uma região de cargas negativas movem-se para uma região de cargas positivas.

Em termos gerais, existem dois tipos de descargas: na nuvem e no solo. Descargas na nuvem originam-se frequentemente dentro das nuvens cumulonimbus e propagam-se dentro (intra-nuvem) ou fora da nuvem em direção a outra nuvem (entre-nuvens) ou ainda fora da nuvem em uma direção qualquer da atmosfera (descarga

para o ar). Descargas no solo tem início dentro da nuvem (nuvem-solo) ou no solo (solo-nuvem). As descargas intra-nuvens são as que ocorrem com mais frequência pois neste caso cargas opostas estão mais próximas do que em qualquer outro tipo de descarga e também pelo fato de que a densidade do ar é menor a medida que a altura vai aumentando, diminuindo assim a capacidade isolante do ar. Mundialmente, as descargas intra-nuvens representam cerca de 70% do número total de descargas, porém este número varia de acordo com a latitude geográfica, sendo em torno de 80 a 90% em regiões próximas ao equador geográfico e em torno de 50 a 60% em localidades de latitudes médias (INPE, 2015).

As descargas atmosféricas entre nuvens e terra são classificadas em função da polaridade da carga que efetivamente é transferida para o solo (positiva ou negativa) e a direção da evolução do canal de descarga (ascendente ou descendente), como é mostrado na Figura 1:

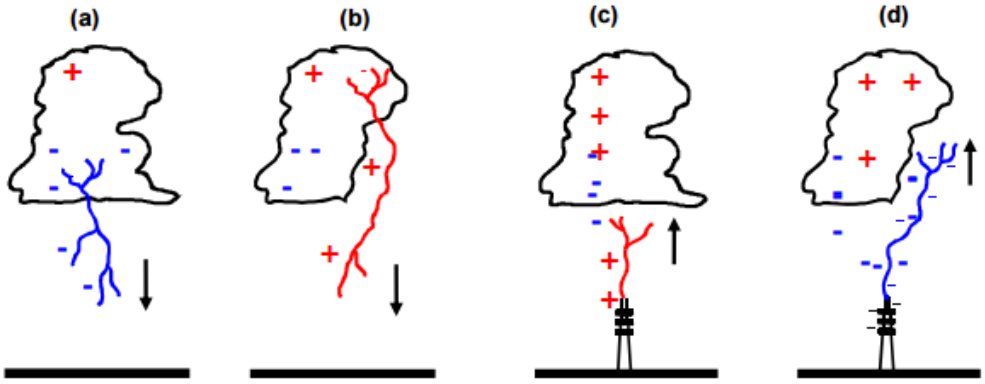


FIGURA 1: Descargas no solo, definidas pela polaridade da carga que é transferida para o solo e a direção do canal (indicada pela seta): (a) negativa descendente, (b) positiva descendente, (c) negativa ascendente e (d) positiva ascendente

FONTE: Adaptado de Heidler *et al.* (2008)

2.1.2 Descargas Nuvem-Solo

A descarga nuvem-solo é o tipo de descarga de maior interesse e estudo pelo fato de atingir diretamente a todos os seres vivos, podendo provocar diversos tipos de destruição. Segundo o INPE (2015), mais de 99% das descargas no solo são do tipo nuvem-solo (descargas solo-nuvem geralmente ocorrem no topo de montanhas ou

estruturas altas). Descargas negativas transportam elétrons de uma região carregada negativamente dentro da nuvem para o solo, já as descargas positivas transferem elétrons do solo para a nuvem. As descargas mais destrutivas são as de polaridade positiva, porém bem menos frequentes do que as de polaridade negativa (globalmente as descargas negativas representam cerca de 90% das descargas nuvem-solo, também podendo este percentual variar de acordo com a localidade (INPE, 2015)).

As descargas duram em torno de um quarto de segundo em média, percorrendo na atmosfera trajetórias com comprimentos desde alguns quilômetros até algumas dezenas de quilômetros. A corrente elétrica flui pelo canal do relâmpago, que é um canal com diâmetro de poucos centímetros onde a temperatura chega a algumas dezenas de milhares de graus e a pressão chega a dezenas de atmosferas. Sendo assim, a corrente elétrica sofre intensas variações desde algumas centenas de àmperes até centenas de quiloàmperes.

Apesar da descarga parecer contínua, geralmente é composta por múltiplas descargas, chamadas de descargas de retorno (ou *return strokes* em inglês), que ocorrem em intervalos de tempo muito curtos. Variações rápidas e lentas de corrente podem acontecer durante o intervalo entre as descargas.

2.1.3 Descargas de Polaridade Negativa Descendente

Uma descarga de polaridade negativa se dá através de fracas descargas na região de cargas negativas dentro da nuvem, em geral em torno de 5 quilômetros, que movem-se até o centro inferior de cargas positivas por um período de cerca de 10 milissegundos denominado período de quebra de rigidez preliminar.

Quebrada a rigidez, o líder escalonado (uma fraca descarga com velocidade em torno de 400.000 km/h, geralmente não visível) se propaga para fora da nuvem em direção ao solo pelo canal do relâmpago. O líder escalonado é dito ser negativo por transportar cargas negativas.

2.1.4 Líder Escalonado

O líder escalonado percorre caminhos sinuosos através de etapas (que duram aproximadamente um microsegundo deslocando-se de 30 a 100 metros), cada qual buscando o trajeto mais fácil para a formação do canal do relâmpago. Durante as etapas, pequenas pausas (cerca de 50 microsegundos) acontecem, sendo que a maior parte da luminosidade é produzida durante as etapas e não nas pausas.

A carga que o líder escalonado transporta ao todo é de dez ou mais coulombs e a corrente média é de algumas centenas de àmperes, com pulsos de ao menos um quiloàmpere correspondentes a cada etapa. Normalmente o líder escalonado ramifica-se por diversos trajetos, porém na maioria das vezes apenas um ramo atinge o solo.

Ao passo que o líder escalonado vai se aproximando do solo (algumas dezenas a pouco mais de uma centena de metros) é gerado um campo elétrico intenso entre o solo e a extremidade do líder, decorrente das cargas elétricas no canal do relâmpago. Este campo tem um potencial elétrico da ordem de 100 milhões de volts e é ele o responsável por exceder a capacidade isolante do ar em um ou mais pontos no solo fazendo com que um ou mais líderes ascendentes positivos, denominados líderes conectantes, surjam do solo difundindo-se de maneira similar ao líder escalonado. Em cerca de um terço dos casos, mais de um líder surge a partir de diferentes pontos no solo.

2.1.5 Descarga de Retorno

Assim que os líderes conectante e escalonado se encontram, as cargas armazenadas no canal movem-se em direção ao solo como uma intensa descarga acompanhada de um forte clarão que se desloca para cima ao longo do canal, iluminando todas as suas ramificações com cerca de um terço da velocidade da luz. Esta descarga recebe o nome de descarga de retorno e produz a maioria da luz visível, durando aproximadamente 100 microsegundos.

Normalmente, o pico de corrente da descarga de retorno é atingido em cerca de 10 microsegundos e decai pela metade em cerca de 100 microsegundos. No período que precede o encontro entre os líderes conectante e escalonado, a corrente aumenta lentamente, passando a subir mais rapidamente e apresentando máxima variação pouco antes de atingir o pico. Atingido seu máximo valor, a corrente começa a diminuir lentamente revelando que menos cargas são depositadas na parte de cima do canal durante o movimento de subida do líder escalonado. Em média, dez coulombs de carga negativa são transferidas ao solo durante uma descarga de retorno.

Cerca de um quinto das descargas negativas são simples, isto é, a descarga termina após a descarga de retorno, entretanto este número pode variar amplamente de acordo com a tempestade. Na maioria dos casos, uma nova descarga chamada de descarga de retorno subsequente se origina decorrente da situação favorável deixada pela primeira descarga de retorno. Entre uma descarga de retorno e outra acontecem pequenas pausas que duram em média de 40 a 90 milissegundos. Para que ela ocorra, cargas dentro da nuvem são transportadas para a região onde se iniciou o líder escalonado. O pico de corrente das descargas de retorno subsequentes tende a ser menor do que a intensidade da primeira descarga de retorno, atingindo seu valor mais rapidamente (em torno de 1 microsegundo) por causa da menor extensão do líder conectante e por durar aproximadamente metade do tempo de uma descarga de retorno. O número de descargas de retorno subsequentes determina a multiplicidade da descarga. Em média, uma descarga negativa possui de 3 a 6 descargas de retorno, sendo que em cerca de 1% dos casos 6 ou mais descargas ocorrem.

2.1.6 Líder Contínuo

O líder contínuo é formado quando as novas cargas transportadas dentro da nuvem alcançam a região do canal formado pela descarga de retorno inicial, abrindo caminho para a descarga de retorno subsequente.

O líder contínuo propaga-se ao longo do canal já ionizado pelo líder escalonado

como um segmento de corrente com um comprimento de 10 a 100 metros, de forma contínua e sem ramificações, durando em média um milissegundo com velocidade média em torno de 40.000 km/h. Em muitos casos, o líder contínuo pode desviar-se ao longo do trajeto seguindo um novo trajeto, em consequência do decaimento do canal inicial ou por causa de ventos fortes, se assimilando a um líder escalonado, sendo chamado líder contínuo-escalonado. Este ocorre mais frequentemente quando o tempo após uma descarga de retorno é superior a 100 milissegundos e nestes casos a velocidade do líder geralmente é menor e a nova descarga de retorno ocorrerá a partir de um líder conectante oriundo de um ponto diferente no solo. Cerca de 30 a 50 % das descargas negativas são desta natureza, denominadas descargas bifurcadas. Casos incomuns ocorrem quando o líder contínuo interrompe abruptamente seu caminho na atmosfera, não gerando uma descarga de retorno subsequente.

Após cerca de 50 microsegundos do início do líder contínuo, quando está a alguns metros de atingir o solo, novamente surge um líder conectante de poucos metros de extensão, tendo-se então a descarga de retorno subsequente. A descarga de retorno subsequente tende a ter velocidade levemente superior do que a velocidade da primeira descarga de retorno.

2.1.7 Descargas de Polaridade Positiva

As descargas de polaridade positiva geralmente iniciam-se a partir de um líder com uma luminosidade não tão intensa quanto a do líder escalonado de uma descarga de polaridade negativa. Este líder se propaga a partir de uma região de cargas positivas dentro da nuvem, apresentando uma luminosidade contínua, e não etapas, mas com variações frequentes de intensidade. Na grande maioria, as descargas positivas apresentam apenas uma descarga de retorno, com intensidade média levemente superior do que as descargas negativas.

2.1.8 Distribuição Estatística do Pico de Corrente

Esta discussão é limitada essencialmente para descargas de retorno negativas descendentes. Levantamentos de dados descargas, no mundo todo, mostram que as variações de pico de corrente (I), correspondente ao valor máximo atingido pela corrente, apresenta uma distribuição lognormal (BERGER; ANDERSON; KRÖNINGER, 1975; BRAGA, 2009):

$$f(I) = \frac{1}{\sqrt{2\pi}\sigma_I I} e^{\frac{-(ln\frac{I}{\mu_I})^2}{2\sigma_I^2}} \tag{1}$$

onde $f(I)$ é a função densidade de probabilidade do pico de corrente da descarga, μ_I, σ_I são, respectivamente, os valores médio e o do desvio padrão, referenciados à primeira componente da descarga.

A probabilidade do pico de corrente de uma descarga (I) exceder um determinado valor I_0 pode ser calculada pela função (P):

$$P(I > I_0) = \frac{1}{1 + (\frac{I_0}{31})^{2.6}} \tag{2}$$

onde:

$P(I > I_0)$: probabilidade do pico de corrente I exceder I_0;

I : pico de corrente da descarga;

I_0 : corrente de descarga de um determinado valor.

A função P representa o complemento da função acumulada de probabilidade $F(I) = P(I \leq I_0)$. Esta distribuição é considerada válida para correntes menores que 200 kA. Seu gráfico é mostrado na Figura 2.

2.2 OCORRÊNCIA DE DESCARGAS NO PLANETA

A cada segundo, cerca de 50 a 100 relâmpagos ocorrem no mundo, o que equivale a cerca de 5 a 10 milhões por dia ou cerca de 1 a 3 bilhões por ano. Embora a maior parte do planeta seja coberta por água, menos de 10% das descargas ocorrem nos oceanos, devido à dificuldade dos oceanos responderem às mudanças de tempera-

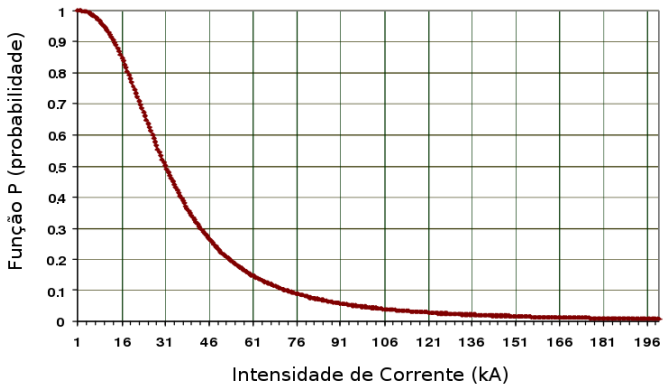


FIGURA 2: Intensidade da corrente $\times$ probabilidade de ocorrência
FONTE: Adaptado de Braga (2009)

tura durante o dia, o relevo e a menor concentração de aerossóis em comparação com a superfície continental. O verão é a estação do ano com maior número de ocorrências de descargas, devido ao maior aquecimento solar, embora ocorram em qualquer período do ano (INPE, 2015).

De um modo geral, sabe-se que as principais regiões de ocorrência de descargas são as regiões sul e central da África, o sul da Ásia, região sul dos Estados Unidos, o Brasil (exceto pela região nordeste), a região norte da Argentina, a ilha de Madagascar, a Indonésia e a região norte da Austrália. A Figura 3 mostra a taxa anual de descargas (por km^{2}ano) na Terra com resolução de $0,5^\circ \times 0,5^\circ$, obtida pela combinação de dois instrumentos de detecção de descargas, o *Optical Transient Detector* (OTD) provendo dados de 1995 a 2000 e o *Lightning Imaging Sensor* (LIS) fornecendo dados de 1998 a 2010 (CECIL; BUECHLER; BLAKESLEE, 2014).

O Brasil é um dos países de maior ocorrência de descargas no mundo devido a sua grande extensão territorial e por estar próximo do equador geográfico. Estima-se que cerca de 50 a 60 milhões de descargas nuvem-solo atinjam o território brasileiro por ano e de 2000 a 2014 foram registradas 1.792 mortes causadas por descargas (INPE, 2015).

A ocorrência de descargas vem aumentando significativamente sobre os centros urbanos em relação às áreas vizinhas, afirma o INPE (2015). Acredita-se que isto

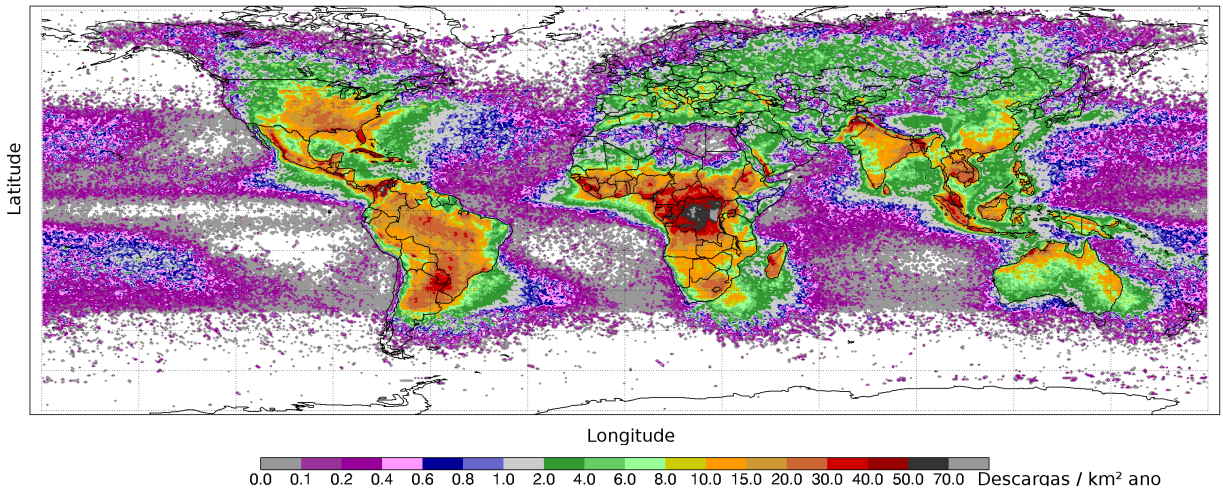


FIGURA 3: Distribuição global de descargas

FONTE: Adaptado de Cecil, Buechler e Blakeslee (2014)

esteja ocorrendo graças ao maior grau de poluição sobre estas regiões e ao aquecimento provocado pela alteração do tipo de solo e a presença de prédios e outros elementos que alteram a temperatura local.

2.3 IMPACTOS NO SETOR ELÉTRICO CAUSADOS POR DESCARGAS

Descargas atmosféricas causam muitos transtornos no setor elétrico, tais como desligamentos de linhas de transmissão, variações de tensão e danos nos equipamentos.

Segundo Silva *et al.* (2010), os prejuízos causados por descargas atmosféricas no setor elétrico são: desligamento parcial ou total de um alimentador, queima de transformadores, danos em pára-raios, rompimento de condutores de alta/baixa tensão, isoladores danificados, banco de capacitores danificados, entre outros. Além disso, descargas atmosféricas são causas de indenizações às seguradoras quanto ao ressarcimento pela queima de equipamentos eletrônicos dos consumidores e também no comprometimento das concessionárias na qualidade da energia fornecida (SHIGA, 2007).

Basicamente, as descargas podem ocasionar surtos na rede elétrica de três mo-

dos: incidindo diretamente nos condutores (descargas diretas), atingindo algum ponto nas proximidades da linha (descargas indiretas) ou afetando diretamente uma edificação (SILVA NETO, 2004).

Entre 1992 e 1997, 26% dos pagamentos de apólices de seguro foram devidos a surtos de origem atmosférica na Alemanha. Na Colômbia entre 1993 e 1994, 30% de aumento foi registrado nos pagamentos de apólices por danos em equipamentos elétricos que chegou a US$ 16 milhões em 1994 (IBÁNEZ et al., (2006) [1] apud SHIGA (2007)).

Cerca de 30% dos desligamentos de energia e prejuízos no setor elétrico avaliados em mais de US$ 1 bilhão por ano são causados por descargas atmosféricas nos Estados Unidos, segundo a *National Lightning Safety Institute*, entidade americana voltada à consultoria e pesquisa na área de proteção contra descargas atmosféricas (LEITE *et al.*, 2009).

No Brasil o prejuízo estimado é de R$ 500 milhões por ano para os setores industrial, telecomunicações e principalmente o setor elétrico, onde cerca de 70% dos desligamentos na transmissão e 40% dos problemas de distribuição de energia são ocasionados por descargas atmosféricas, segundo o INPE (2015). As descargas atmosféricas representam grande impacto para o sistema de distribuição de energia elétrica brasileiro devido à configuração predominantemente aérea das linhas (cerca de 99%), significando que estão completamente expostas às condições climáticas.

Aproximadamente 47% de falhas em transformadores de distribuição da Companhia Energética de Minas Gerais (CEMIG) são causados por descargas atmosféricas, acarretando prejuízos de US$ 1,83 milhões por ano, segundo Couto, Duarte e Soares (1995).

“Grande parte das interrupções no fornecimento de energia elétrica no Estado do Paraná têm como causa descargas atmosféricas” (COPEL, 2014).

[1] IBÁNEZ, H. F.; AVENDAÑO, C. A.; ORTIZ, H. E. Correlação entre Nível Ceuráunico e Danos em Aparelhos Eletroeletrônicos. **Eletricidade Moderna**, n. 390, p. 72-79, 2006

Análises sobre a relação de descargas atmosféricas e sistemas elétricos podem ser encontrados em: Chowdhuri (1989), Cummins, Krider e Malone (1998), Diendorfer e Schulz (2003), King (2003), Visacro, Dias e Mesquita (2005), Nucci (2010).

2.3.1 O Setor Elétrico e o Pico de Corrente das Descargas

Sabe-se que o pico de corrente da descarga é uma variável significativa no potencial destrutivo do fenômeno natural em questão. Especificamente falando sobre o setor elétrico, descargas com alto pico de corrente tem maior probabilidade de ocasionar falhas em linhas de transmissão de energia (WESTINGHOUSE, 1964; LEITE *et al.*, 2009). Esta conclusão é fundamentada pelo fato de que a incidência direta de uma descarga sobre um condutor de uma linha de energia causa pulsos de alta voltagem, que se propagam como ondas que viajam em ambas as direções a partir do ponto atingido. A crista do pulso pode ser calculada como (VIJAYARAGHAVAN; BROWN; BARNES, 2004):

$$V = I \times Z \tag{3}$$

onde V é a voltagem da crista, I é o pico de corrente da descarga e Z é a impedância vista pelo pulso ao longo da direção de viagem. A impedância Z é um valor característico da linha de energia e, portanto, é a intensidade do pico de corrente que vai determinar a voltagem imposta a linha de energia. Quanto maior for o pico de corrente, maior será a tensão da voltagem, e se o nível de isolamento da linha é inferior a tensão imposta, poderá haver rompimento e falha.

Mais estudos sobre a relação do pico de corrente de descargas e ocorrências de interrupções em linhas de energia podem ser encontrados em Diendorfer e Schulz (2003), Diendorfer e Pistauer (2010).

Assim, é de suma importância conhecer o comportamento desta variável que, para o setor elétrico, pode representar benefícios significantes. O sistema proposto nesta pesquisa é capaz de monitorar e fazer previsões do pico de corrente das descargas de tempestades elétricas.

2.4 MATERIAIS E FERRAMENTAS UTILIZADOS NA PESQUISA

2.4.1 Região de Estudo

A fim de garantir melhor qualidade de energia para seus consumidores com o mínimo de interrupções e satisfazendo os requisitos de qualidade, as concessionárias vem investindo em programas de pesquisa e desenvolvimento na prevenção de perturbações causadas por descargas atmosféricas.

Conhecer as características das tempestades elétricas que atuam em regiões que abrangem linhas de transmissão de energia pode significar grande avanço para o setor, pois medidas operacionais que minimizem riscos de desligamentos podem ser tomadas. Esse foi o incentivo para estudar a região da linha de transmissão de energia mais importante do Brasil: a linha LT 765 kV, conhecida como Linhão de Itaipu. Esse sistema leva a energia produzida na usina hidrelétrica de Itaipu para a proximidade do centro de consumo da região Sudeste do Brasil. O sistema é composto de três linhas de transmissão (cada uma com extensão de aproximadamente 900 quilômetros e cerca de 2 mil torres de transmissão) e inicia-se na subestação de Foz do Iguaçu (PR), passa pelas subestações de Ivaiporã (PR) e a de Itaberá (SP) e termina em Tijuco Preto (SP), na região metropolitana de São Paulo (FIGURA 4). Entre cada subestação, a linha se divide em três circuitos, contabilizando ao todo nove circuitos ao longo da linha. Iniciou sua operação em 1986 e, até hoje, é o sistema de transmissão de tensão mais elevada existente no Brasil (ITAIPU, 2013).

2.4.2 Dados de Descargas

Os dados de descargas atmosféricas utilizados neste trabalho são provenientes da Rede Integrada Nacional de Detecção de Descargas Atmosféricas (RINDAT) mantidos pelo Sistema Meteorológico do Paraná (SIMEPAR) em cooperação com Furnas, CEMIG e INPE. Desde o ano de 1999 em diante, dados de descargas atmosféricas são armazenados com as seguintes informações: posição geográfica (latitude/longitude),

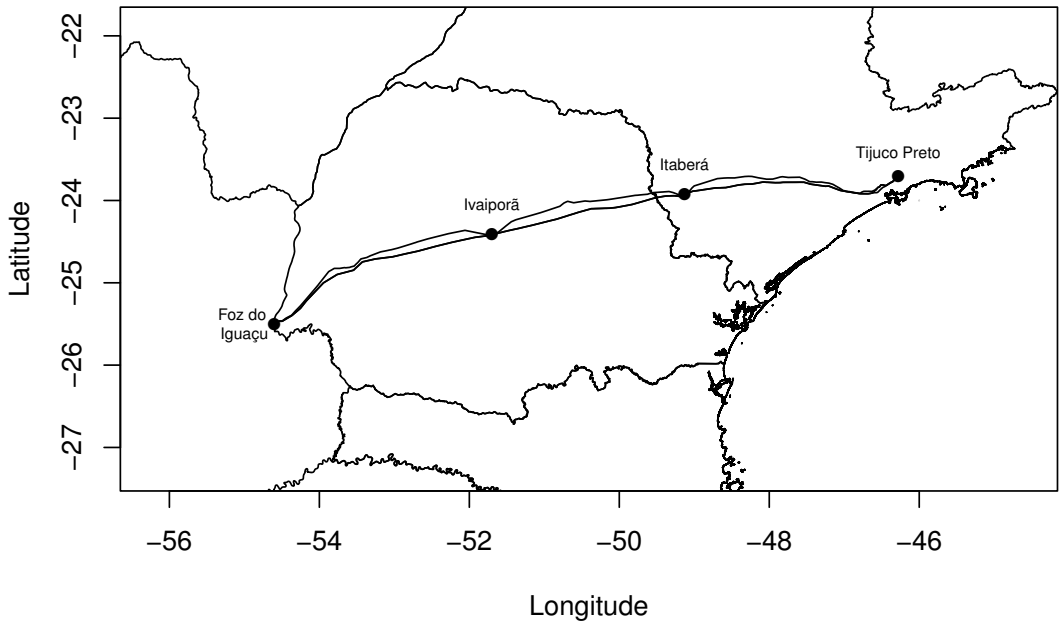


FIGURA 4: Região do Brasil que abrange a Linha LT 765 kV
FONTE: A autora (2015)

tempo de ocorrência (hora, minuto, segundo e milissegundo em UTC – *Universal Time Coordinated*), polaridade (positiva ou negativa) e intensidade do pico de corrente elétrica (em kA), além de outras características elétricas da forma de onda.

Os dados processados contemplam apenas descargas com pico de corrente igual ou superior a 10 kA em módulo, uma vez que intensidades menores que esta indicam prováveis descargas intra-nuvens e é sugerido que sejam filtradas para eliminar ruídos nas informações (CUMMINS et al., 1998 [2] apud BENETI, 2012).

2.4.3 Ferramenta Computacional

O *software* adotado para análise e processamento de dados é o R (R Core Team, 2012). É um ambiente de domínio público, livre e que apresenta código fonte aberto, podendo ser modificado ou implementado com novos procedimentos desenvolvidos pelos próprios usuários. É uma linguagem de programação simples, de fácil elabora-

[2]CUMMINS, K. L.; KRIDER, E. P.; MALONE, M. D. The U.S. National Lightning Detection Network and applications of cloud-to-ground lightning data by electric power utilities. **IEEE Transactions on Electromagnetic Compatibility**, v. 40, p. 465, 1998

ção de gráficos, amplamente utilizado nas mais diversas áreas científicas, especialmente no campo estatístico.

3 ANÁLISE DE AGRUPAMENTOS DE DADOS

A clusterização ou agrupamento de dados é uma técnica que contribui para a interpretação de características primordiais de grandes conjuntos de informações, geralmente incompreensíveis na forma como são apresentados de maneira bruta. Basicamente, o conceito de clusterização é agrupar dados similares no mesmo grupo, também chamado de *cluster*. Assim, dados de um mesmo *cluster* tem mais características em comum entre si do que com dados de outros *clusters* (BOSCARIOLI, 2008). Por isso, o agrupamento de dados é feito com base na similaridade entre os dados.

É fundamental distinguir classificação de clusterização. O primeiro tem por objetivo alocar um novo dado à grupos já definidos, já o segundo não conhece *a priori* os grupos e sua tarefa é encontrá-los (OLIVEIRA, 2008; CAVALCANTI JÚNIOR, 2006). Sendo assim, clusterização é um processo de aprendizado não supervisionado, uma vez que não existem exemplos pré-definidos que evidenciem que algum tipo de relação deva existir entre os dados. Segundo Boscarioli (2008), um sistema de aprendizado não supervisionado utiliza análise de redundância da informação recebida para adquirir conhecimento, pois não existe, ou não se assume, um conjunto de dados para testes de classificação. Essa redundância pode ser obtida pelo cálculo de propriedades estatísticas do conjunto de informações ou analisando a forma de como os dados podem estar agrupados.

Um algoritmo de clusterização deve levar em conta diversos aspectos, tais como: a forma com que os dados estão sendo representados, como medir a similaridade entre dados e entre *clusters* e como estimar a qualidade do resultado obtido pelo método. A definição de diferentes algoritmos de clusterização se dá pela maneira como estes aspectos são abordados juntamente com a escolha dos parâmetros iniciais.

3.1 APLICAÇÕES PRÁTICAS DE CLUSTERIZAÇÃO

A análise de *cluster*, tem por objetivo auxiliar um usuário a compreender o agrupamento ou estrutura natural em um conjunto de informações. A metodologia de clusterização tem sido bastante utilizada em diversas aplicações como reconhecimento de padrões, análise de dados, pesquisa de mercado, processamento de imagens, aprendizado de máquinas, entre outras. Alguns trabalhos que utilizam clusterização de dados em aplicações na área ambiental serão brevemente relatados a seguir.

Zagouras *et al.* (2013) utilizam clusterização de dados de nuvens para encontrar a melhor localização para a instalação de instrumentos que captam energia solar na Grécia. O resultado do agrupamento revela que a variabilidade da radiação solar da superfície devido à nebulosidade sobre a região de aplicação poderia ser monitorada adequadamente com a implantação de 22 instrumentos terrestres. A representatividade espacial dos locais propostos também é avaliada. O número de estações propostas pode ser considerado como a base para construir a climatologia da superfície de radiação solar sobre a Grécia.

Freitas *et al.* (2013) utilizam alguns métodos de clusterização para identificar regiões homogêneas com 30 anos de dados de postos pluviométricos para os índices climáticos de aridez, umidade e hídrico, obtidos por meio do balanço hídrico climatológico para o estado da Paraíba/BR. Com a escolha do melhor método de clusterização para cada cenário e com o auxílio do dendrograma, 5 regiões foram identificadas pelo índice de aridez, 4 pelo de umidade e 5 pelo hídrico. A análise de *cluster* evidenciou a existência de diferentes regiões da Paraíba com maior ou menor potencialidade hídrica para o cultivo de culturas de subsistência.

Bonato (2014) aplica a técnica de clusterização *K-Means++* para classificar células de tempestades em convectiva ou estratiforme, utilizando dados de refletividade estimados por um radar meteorológico.

Woolford e Braun (2006) estudam o relacionamento de incêndios florestais e des-

cargas atmosféricas em Ontário no Canadá por meio de clusterização de dados no espaço-tempo. Eles utilizam uma técnica chamada *Convergent Data Sharpening* para identificação dos centros das tempestades, que se baseia no deslocamento de dados para mais perto de seu modo local a cada iteração, fazendo com que os dados fiquem mais agrupados. Os autores propõem iterar o algoritmo, mostrando que os dados irão convergir tanto para modos locais ou globais. A ideia da aplicação é representar a massa de dados de descargas pelo centro da tempestade elétrica que as conduz e então associá-lo ou não a focos de incêndios na província.

3.2 MEDIDAS DE PROXIMIDADE

3.2.1 Proximidade entre Dados

Para que dados possam ser agrupados, é imprescindível alguma medida quantitativa indicando o quão próximos ou distantes os dados estão uns dos outros. Essa proximidade pode representar a similaridade, dissimilaridade ou distância entre os dados. O que se faz então é reunir esse conjunto de informações contendo a proximidade entre os dados e dispô-lo em uma matriz, que recebe o nome de matriz de similaridade quando o conjunto de informações é a similaridade e matriz de dissimilaridade quando o conjunto é a dissimilaridade. Quanto maior a similaridade (ou menor a dissimilaridade) entre os dados, mais próximos estes dados encontram-se.

A medida de similaridade utilizada para representar distância entre dados depende muito do tipo de dado que se está trabalhando e a escala adotada. A escala indica o grau de importância de uma característica do dado em relação as demais (BOSCARIOLI, 2008). Quando as características de um dado estão em escalas muito distintas, o que costuma-se fazer é uma padronização dos dados, que consiste em converter as características originais em valores sem unidade de medida, porém a padronização carrega a desvantagem de reduzir todas as variáveis ao mesmo grau de agrupabilidade. As padronizações mais usadas frequentemente são as padronizações Z e Min-Max (SILVA; SANTOS, 2007).

A medida de dissimilaridade mais utilizada é a distância de Minkowski, que mede a separação entre dois dados contínuos x_i e x_j, dada pela equação:

$$d(x_i, x_j) = \sqrt[p]{\sum_{k=1}^{m} (|x_{ik} - x_{j_k}|)^p}, \quad p \geq 1 \tag{4}$$

onde $x_i = (x_{i1}, x_{i2}, \ldots, x_{im})$. Valores diferentes do parâmetro p resultam em distâncias distintas: $p = 1$ é a distância de Manhattan, $p = 2$ é a distância euclidiana e $p \to \infty$ é a chamada distância *sup*. Para distâncias entre dados não contínuos, consultar trabalho de Lopes (2006).

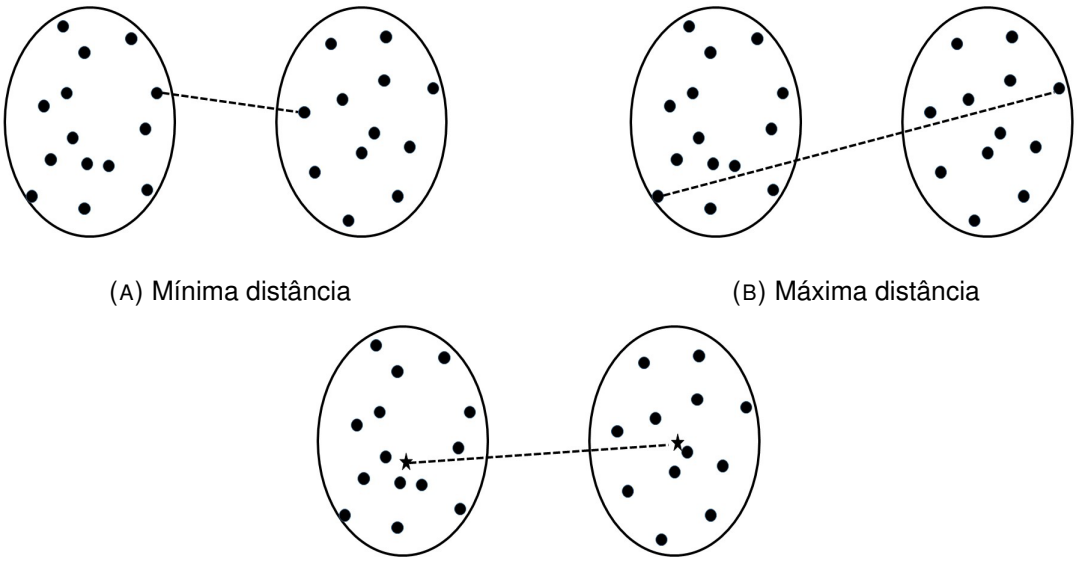
3.2.2 Proximidade entre *Clusters*

Além de medir a proximidade entre elementos, alguns algoritmos de clusterização requerem medidas entre *clusters* (por exemplo, para unir grupos similares). Uma maneira de medir essa proximidade é calcular a distância entre todos os pares de pontos dos *clusters*, em que cada ponto pertence a um *cluster* distinto. Se a distância mínima entre todos os pares de pontos é escolhida, tem-se a mínima distância (*single linkage*); se for escolhida a distância máxima entre todos os pares de dados, tem-se a máxima distância (*complete linkage*); e se for escolhida a distância entre os centroides dos *clusters*, tem-se a distância pelo centroide (*average linkage*) (BOSCARIOLI, 2008). A Figura 5 ilustra as três distâncias mencionadas.

A maneira de se medir a distância entre *clusters* está sujeito à aplicação e ao formato dos grupos. Segundo Baeza-Yates e Frakes [1] (1992) apud Boscarioli (2008) a máxima distância é melhor para *clusters* mais compactos, e de acordo com Rocha e Lanchi [2] (2005) apud Boscarioli (2008) a mínima distância gera *clusters* mais alongados.

[1] BAEZA-YATES, R. A.; FRAKES, W. B. **Introduction to data structures and algorithms related to information retrieval**. San Diego, CA, USA: Prentice-Hall, p. 13–27. 1992.

[2] ROCHA, H. V.; LANCHI, R. L. **Aspectos Básicos de Clustering: Conceitos e Técnicas**. Relatório Técnico IC-05-003, UNICAMP - Universidade Estadual de Campinas, Campinas, fevereiro 2005.



(A) Mínima distância (B) Máxima distância

(C) Centroide

FIGURA 5: Ilustração de medidas de proximidade entre *clusters*, pelo método da (A) mínima distância, (B) máxima distância e (C) centroide

FONTE: Adaptado de Boscarioli (2008)

3.3 TÉCNICAS DE CLUSTERIZAÇÃO DE DADOS

O problema de clusterização pode ser tratado segundo diferentes abordagens, entre elas, o tratamento convencional chamado *hard*, na qual cada objeto deve ser classificado único e totalmente em um determinado *cluster*, e a abordagem *fuzzy*, mais flexível, na qual um objeto pode ser classificado em vários grupos, com diferentes graus de pertinência a cada um deles.

3.3.1 Clusterização do tipo *Hard*

Como já mencionado, clusterização *hard* aloca um objeto exclusivamente em um único grupo, ou seja, a noção de pertinência é categórica: objetos pertencem ou não pertencem a um dado agrupamento.

Existem vários algoritmos para clusterização do tipo *hard*, que usam maneiras diferentes para identificar e representar grupos, dependendo do tipo de dado, da aplicação, entre outros aspectos. A seguir serão brevemente descritas algumas das várias

categorias de algoritmos de agrupamento da categoria *hard*.

Algoritmos de Clusterização por Particionamento

Os algoritmos de clusterização baseados em particionamento dividem uma única vez os dados em um número determinado de *clusters*. Particionar uma única vez pode ser vantajoso quando se tem muitos dados e o armazenamento e processamento de todas as possibilidades de divisões da clusterização hierárquica (descrita adiante) torna-se computacionalmente cara.

Essa classe de método tem por propósito maximizar a similaridade entre elementos de um mesmo *cluster* e minimizar a similaridade entre elementos de *clusters* diferentes, pela otimização de uma função objetivo. Esta função pode representar critérios diferentes a serem otimizados: um critério global usa um ponto representativo de cada *cluster* e agrupa os demais de acordo com a similaridade dos dados com este ponto representativo, já um critério local organiza os grupos pelas informações estruturais dos dados, como por exemplo atribuir um elemento e seus vizinhos mais próximos a um mesmo *cluster* (OLIVEIRA, 2008).

Os algoritmos de particionamento mais utilizados são o *k-means* e o *k-medoids* (SILVA; SANTOS, 2007; CAVALCANTI JÚNIOR, 2006; BORA; GUPTA, 2014), dois algoritmos que se diferenciam pelo tipo de representatividade utilizada para os *clusters*: no algoritmo *k-means* o ponto que representa os demais elementos do *cluster* é o seu centroide ou média, enquanto que o *k-medoids* faz uso do medoide (objeto para o qual a dissimilaridade média de todos os objetos no *cluster* é mínima). A técnica *k-means* é bastante sensível a ruídos pois um elemento com um valor extremamente grande pode distorcer a distribuição dos dados, já o algoritmo *k-medoids* busca reduzir essa sensibilidade a ruídos já que utiliza o medoide como ponto de referência de um *cluster*. O método *k-means* é sensível a partição inicial e não é adequado para descobrir *clusters* com formas não convexas e tamanhos muito diferentes (CARLANTONIO, 2001).

Algoritmos de Clusterização por Hierarquia

Nesta classe de algoritmo, os dados são agrupados ou separados conforme a proximidade entre eles. A decomposição hierárquica é representada por um dendrograma, uma espécie de árvore que mostra as decorrentes divisões ou uniões dos dados nos *clusters*. As folhas da árvore simbolizam os dados e conforme a árvore vai crescendo, os dados vão se agrupando para formar grupos maiores, até que todos sejam unidos em um único *cluster*, representado pela raiz. O dendrograma pode ser construído de duas maneiras (CARLANTONIO, 2001):

- Aglomerativo (*bottom-up*): parte-se das folhas superiores para a raiz, isto é, inicialmente os dados representam *clusters* unitários e a cada iteração, os dois *clusters* mais similares são unidos, até que no final, reste apenas um único *cluster* contendo todos os dados;
- Divisivo (*top-down*): parte-se da raiz para as folhas. Todos os dados começam aglomerados em um único *cluster* e a cada etapa um *cluster* é selecionado e dividido em dois *clusters*. No final, tem-se tantos *clusters* quanto o número de dados originais.

Os métodos aglomerativos hierárquicos requerem o cálculo da matriz de dissimilaridades entre os grupos, visto que inicialmente cada elemento é um *cluster*. Então o agrupamento entre elementos, elementos e grupos ou entre grupos se dá por meio da adoção de algum critério em relação a distância. Se a mínima distância é adotada, a técnica recebe o nome de Método de Ligação Simples; se a máxima distância é utilizada, é chamado de Método de Ligação Completa; se a distância média é usada (ponto central do grupo), recebe o nome de Método do Centroide (BOSCARIOLI, 2008; QUINTAL, 2006). Após cada agrupamento, recalcula-se a matriz de dissimilaridade de acordo com os novos grupos formados e repete-se o processo até que todos os elementos estejam em um único grupo. O algoritmo AGNES (*Agglomerative Nesting*) possivelmente seja o método aglomerativo hierárquico mais conhecido (CARLANTO-

NIO, 2001).

Já os métodos divisivos hierárquicos fazem o caminho contrário dos métodos aglomerativos. O algoritmo DIANA (*Divisive Analysis*) inicia com todos os elementos em um único *cluster*, e em cada passo, o *cluster* com maior diâmetro é dividido em dois *clusters*. O diâmetro de um cluster é definido como a dissimilaridade máxima entre todos os elementos dentro de um *cluster*. Recalcula-se a dissimilaridade entre os *clusters* e repete-se o processo até que cada novo *cluster* contenha somente um elemento simples (CARLANTONIO, 2001).

Embora bastante simples, os métodos hierárquicos não são capazes de efetuar ajustes uma vez que uma união ou divisão tenha sido feita. Esse fato pode comprometer o resultado final visto que, uma vez que um conjunto de dados é unido ou particionado, a próxima iteração será processada com base nos grupos recém-formados. Outra desvantagem desse tipo de método é a sensibilidade com respeito a ruídos e também ao alto custo computacional quando grandes conjuntos de dados são utilizados (BOSCARIOLI, 2008).

Diferentemente dos algoritmos de particionamento, os métodos hierárquicos não necessitam do número de *clusters* como parâmetro de entrada, porém uma condição de término deve ser adotada indicando quando o processo de divisão ou união de grupos deve acabar (CARLANTONIO, 2001).

Algoritmos de Clusterização por Densidade

São baseados no conceito de densidade em que os *clusters* vão crescendo de acordo com a densidade da vizinhança dos objetos ou por meio de uma função de densidade. Sendo assim, os *clusters* representam regiões com alta densidade separados por regiões de baixa densidade. Portanto, para cada dado de um *cluster*, a vizinhança deve conter um número mínimo de objetos (CARLANTONIO, 2001; BOSCARIOLI, 2008).

Essa classe de métodos tem a vantagem de formar grupos de formas arbitrárias

capaz de filtrar ruídos. Outro ponto positivo é que o método não necessita do número de *clusters* como parâmetro inicial.

DBSCAN (*Density Based Spatial Clustering of Applications with Noise*) é um algoritmo que utiliza a identificação de *clusters* em áreas de alta densidade. A ideia do método é que cada elemento de um *cluster* tenha uma vizinhança com um número mínimo de pontos. Se *Eps* é o raio máximo da vizinhança formada por um número mínimo de pontos denotado por *MinPts*, a vizinhança de um ponto x_i é definida por $N_{Eps}(x_i) = \{x_j \in X | d(x_i, x_j) \leq Eps\}$, onde $d(x_i, x_j)$ é uma medida de distância a qual define o formato da vizinhança. Após a identificação das regiões densas, os pontos são classificados em relação a sua posição: x_i é um ponto central se $|N_{Eps}(x_i)| \geq MinPts$; x_i é ponto de borda se pertence a vizinhança de um ponto central x_j; caso contrário é dito ponto ruidoso. As entradas do algoritmo são: *Eps*, *MinPts* e o conjunto a ser clusterizado X. Um *cluster* é caracterizado por um conjunto de pontos densamente conectados (ESTER *et al.*, 1996). Dois pontos centrais são agrupados se a distância entre eles é menor que *Eps*, pontos de borda são colocados no mesmo *cluster* que os pontos centrais e pontos ruidosos são descartados da classificação por não pertencerem a nenhum *cluster*. Maiores detalhes sobre o algoritmo DBSCAN podem ser encontrados em Ester *et al.* (1996) e Carlantonio (2001).

O algoritmo OPTICS (*Ordering Points To Identify the Clustering Structure*) (ANKERST *et al.*, 1999) é uma extensão do DBSCAN para que diversos valores de distância *Eps* sejam processados simultaneamente, assim grupos com diferentes densidades são formados ao mesmo tempo. Para produzir um resultado consistente, os objetos são selecionados de forma que são alcançáveis por densidade com o menor valor de *Eps*, para garantir que grupos de alta densidade sejam terminados primeiro. O algoritmo trabalha para um número infinito de Eps_i menores ou iguais a *Eps*. A saída é uma ordenação da base de dados que possibilita extrair agrupamentos baseados em densidade para infinitas configurações de parâmetros a um custo computacional reduzido.

Algoritmos de Clusterização por Grades

Esta classe de método utiliza a estrutura de dados em grade de multi-resolução, dividindo o espaço dos dados em um número finito de células, que configuram a estrutura de grade em que todos os procedimentos de clusterização são executados (CARLANTONIO, 2001; BOSCARIOLI, 2008). Esta metodologia divide o espaço dos dados em células disjuntas na grade e há grande chance de que todos os dados que estejam em um mesmo grupo também estejam na mesma célula da grade. Assim, os dados pertencentes a uma mesma célula podem ser representados e tratados como um único dado (QIU; ZHANG; SHEN, 2005).

O método STING (*Statistical Information Grid*) divide a área espacial em células retangulares, onde existem diversos níveis correspondentes a resolução destas células, formando uma estrutura hierárquica: cada célula do nível mais alto é dividida para originar um número de células no próximo nível mais baixo. Informações estatísticas de cada célula são calculadas e armazenadas de antemão. Parâmetros de células de nível mais elevado podem ser facilmente calculados a partir de parâmetros da célula de nível inferior (como média, variância e tipo de distribuição). As informações estatísticas são úteis para processos de responder consultas. Inicialmente, é definida a camada da estrutura hierárquica onde o processo de consulta começará. Para todas as células da camada atual, o intervalo de confiança é calculado, revelando a importância da célula para a dada consulta. Células com menor importância são retiradas da consideração adicional. Apenas células importantes restantes são avaliadas no processamento do próximo nível mais baixo. Repete-se este processo até que a camada mais baixa seja atingida. Se a especificação da consulta é encontrada, as regiões das células importantes que satisfazem a consulta são retornadas, caso contrário os dados que caem nas células importantes são retomados e processados até que eles satisfaçam os requisitos da consulta (CARLANTONIO, 2001).

O método CLIQUE (*Clustering In Quest*) é adequado para agrupar dados de alta dimensão em ampla base de dados, combinando métodos de clusterização baseados

em grade e em densidade. Multidimensionalmente, um grande conjunto de pontos ocupam não uniformemente o espaço dos dados, assim o CLIQUE identifica áreas (unidades) densas e esparsas no espaço, encontrando padrões globalmente distribuídos do conjunto de dados. Se a parcela do total de pontos de dados contidos em uma unidade excede um parâmetro de entrada do método, esta unidade é dita densa. Desta forma, um *cluster* é definido como um conjunto máximo de unidades densas conectadas (CARLANTONIO, 2001).

Algoritmos de Clusterização por Modelos

Os algoritmos de agrupamentos baseados em modelos tem por meta otimizar o ajuste entre os dados de entrada e algum modelo matemático, isto é, um modelo pressuposto é elaborado para cada *cluster* e encontra-se o melhor ajuste do dado ao modelo. Uma função densidade que dita a distribuição espacial dos dados pode ser usada para localizar grupos. Estes métodos geralmente são baseados na teoria de que os dados são gerados por uma mistura de distribuições de probabilidade subjacentes (BOSCARIOLI, 2008). Dentro desta classe de clusterização, há duas subclasses: *conceitual* e *rede neural*.

- Conceitual: esta abordagem é uma forma de agrupamento em aprendizagem de máquina que, dado um grupo de objetos não rotulados, produz um projeto de classificação sobre eles. É um processo em duas fases: primeiro é realizada divisão e depois é feita uma caracterização dos dados, gerando conceitos e descrições dos grupos.

O método COBWEB é um exemplo de método conceitual. Consiste na criação de uma árvore de classificação guiada pelo uso de uma medida de avaliação estatística chamada de utilidade categórica, isto é, o aumento no número esperado de valores de atributos que podem ser corretamente previstos em um determinado particionamento sobre o número esperado de previsões corretas sem dado conhecimento (CARLAN-

TONIO, 2001). Inicialmente, o método COBWEB incrementalmente atribui os objetos em um árvore de classificação, descendo a árvore por meio de um trajeto apropriado, atualizando as quantidades ao longo do caminho, na procura do melhor nó para classificar o objeto. Esta decisão é baseada na localização temporária do objeto em cada nó e pelo cálculo da utilidade categórica da partição resultante, onde um bom nó para o objeto é indicado pela colocação que resulta na mais alta categoria. O método também calcula a utilidade categórica da partição que resultaria se um novo nó fosse criado para o objeto, o qual é então colocado em uma classe já existente ou é criada uma nova classe para ele, fundamentado no particionamento com o mais alto valor de utilidade categórico.

Outro método conceitual é o *Mclust*: os dados x são vistos como provenientes de uma mistura de G funções densidades $f(x) = \sum_{k=1}^{G} \alpha_k f_k(x)$, em que f_k representa a função densidade de probabilidade dos dados em um grupo k e α_k é a probabilidade que um dado apresenta na *k-ésima* componente da mistura ($0 < \alpha_k < 1$). Cada componente $(1, \ldots, G)$ é modelada por uma distribuição Gaussiana, que é caracterizada pela média μ_k e a matriz de covariância Σ_k, e tem a função densidade de probabilidade:

$$\phi(x_i; \mu_k; \Sigma_k) = \frac{exp\{-\frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1}(x_i - \mu_k)\}}{\sqrt{det(2\pi\Sigma_k)}} \tag{5}$$

Os dados gerados por misturas de densidades normais multivariadas são caracterizados por *clusters* com centro em μ_k, com aumento da densidade nos pontos perto da média. As superfícies correspondentes de uma densidade constante são elipses. Características geométricas dos *clusters* (tais como formato, orientação e volume) são definidas a partir das covariâncias Σ_k (FRALEY; RAFTERY, 2007).

- Rede Neural: Representa cada grupo por meio de um modelo, que não necessariamente corresponde a um dos objetos do grupo. Por meio de alguma medida de distância, novos objetos são atribuídos ao grupo cujo modelo lhe seja mais próximo.

A auto-organização de mapas é uma abordagem de rede neural para clusterização, sendo o SOM o método mais clássico. Consiste em uma rede neural de treinamento não supervisionado, que estrutura topologicamente as unidades de seus agrupamentos. O modelo de padrão de entrada associado a cada agrupamento é representado pelo vetor de peso para cada neurônio (*cluster*, neste caso). Na etapa de treinamento do SOM, o neurônio cujos pesos são mais próximos do padrão de entrada é chamado de neurônio vencedor. Uma vizinhança, em relação a topologia do mapa, é criada para o neurônio vencedor e seus pesos são atualizados, ainda que os vizinhos do neurônio vencedor não necessariamente tenham pesos similares ao padrão de entrada (KOHONEN, 2001; SIQUEIRA, 2005).

Além dos métodos citados, existem muitos outros algoritmos de clusterização na literatura. Em seu trabalho, Boscarioli (2008) fez um apanhado de algoritmos das classes de clusterização descritas anteriormente. Estes algoritmos estão expostos na Tabela 1.

3.3.2 Clusterização do tipo *Fuzzy*

A clusterização do tipo *hard* pode ser inadequada quando há pontos que estão igualmente distantes de dois ou mais grupos, exigindo a designação completa para algum destes grupos, embora exista chance igualitária do ponto pertencer aos demais grupos. A clusterização do tipo *fuzzy* é mais flexível no sentido de que um dado pode pertencer a mais do que um agrupamento ao mesmo tempo, com graus de pertinência ou associação. Graus de pertinência podem expressar o quanto de certeza ou incerteza o dado foi designado ao *cluster* correto. Cada conjunto *fuzzy* é caracterizado pela sua função de pertinência que é uma curva que define o grau de posse (valor entre 0 e 1) de cada ponto, onde geralmente as funções de pertinência mais utilizadas são a triangular, trapezoidal e gaussiana (BIONDI NETO *et al.*, 2006).

Os métodos baseados em clusterização *fuzzy* são portanto mais realistas, uma vez que as fronteiras entre os grupos são de fato muito mal delineadas (DÖRING;

TABELA 1: Exemplos de algoritmos de clusterização, divididos em categorias

Classe		Exemplos de Algoritmos
Particionamento		*k-means*
		k-medoids
		CLARA
		CLARANS
		CLUSTER
		PAM
Hierárquico		AGNES
		DIANA
		CURE
		CHAMELEON
		BIRCH
		ROCK
		CLINK
Densidade		DBSCAN
		OPTICS
		DENCLUE
		GDBSCAN
		DBCLASD
Grade		STING
		CLIQUE
		WAVECLUSTER
		BANG-CLUSTERING
		GRIDCLUST
Modelo	Conceitual	COWEB
		CLASSIT
		AUTOCLASS
		SNOB
		MCLUST
	Rede Neural	SOM
		ART

FONTE: Adaptado de Boscarioli (2008)

LESOT; KRUSE, 2006). Uma desvantagem dos algoritmos *fuzzy* em relação a alguns algoritmos *hard* é que é necessário o conhecimento do número de grupos.

Segundo Döring, Lesot e Kruse (2006), tem-se a seguinte definição: Seja $X = \{x_1, x_2, \ldots, x_n\}$ o conjunto de dados com n elementos e seja k o número de *clusters* $(1 < k < n)$ representado pelos conjuntos *fuzzy* μ_{Γ_i}, $(i = 1, \ldots, k)$. Então $U = (u_{i,j}) = (\mu_{\Gamma_i}(x_j))$ é a clusterização *fuzzy probabilística* de X se $\sum_{j=1}^{n} u_{ij} > 0, \quad \forall i \in \{1, \ldots, k\}$ e $\sum_{i=1}^{k} u_{ij} = 1, \quad \forall j \in \{1, \ldots, n\}$.

O termo $u_{ij} \in [0;1]$ é interpretado como o grau de pertinência do dado x_j ao *cluster* Γ_i em relação a todos os outros *clusters*. Esta definição garante que nenhum *cluster* é vazio e também assegura que a soma dos graus de pertinência para cada dado é 1. Assim cada dado recebe o mesmo peso em comparação com todos os outros dados e, portanto, que todos os dados estão (igualmente) incluídos na partição dos *clusters*. Estas condições garantem que nenhum *cluster* pode conter a adesão plena de todos os dados.

Ainda segundo Döring, Lesot e Kruse (2006), se a condição $\sum_{i=1}^{k} u_{ij} = 1, \quad \forall j \in \{1,\ldots,n\}$ é removida, tem-se uma clusterização *fuzzy possibilística* e $u_{ij} \in [0;1]$ é interpretado como o grau de representatividade do dado x_j ao *cluster* Γ_i.

O mais popular algoritmo de clusterização dessa classe é o *fuzzy c-means*, melhor que o *k-means* (*hard*) para evitar estagnação em mínimos locais (CAVALCANTI JÚNIOR, 2006). É um método de clusterização não hierárquico que proporciona uma partição *fuzzy* de um conjunto de dados em *c clusters*, por meio da minimização de uma função objetivo que mede a adequação entre os dados e os *clusters*. Maiores detalhes sobre o algoritmo *fuzzy c-means* podem ser encontrados em Babuska (1998), Döring, Lesot e Kruse (2006), Bora e Gupta (2014).

3.4 ALGORITMO DE CLUSTERIZAÇÃO UTILIZADO NA PESQUISA

Nesta seção será descrito o algoritmo de clusterização usado nesta pesquisa. É importante ressaltar que existem muitos algoritmos de clusterização pois não há uma técnica de agrupamento universal, capaz de revelar toda a variedade de estruturas que podem estar presentes em conjuntos de dados dos mais diversos tipos e configurações. Para saber qual é a melhor técnica de agrupamento a se utilizar, primeiramente deve-se conhecer a fundo a natureza dos dados, suas propriedades intrínsecas e o que esperar de uma clusterização.

A clusterização de dados na pesquisa terá a finalidade de agrupar descargas atmosféricas próximas no tempo e no espaço, a fim de que descargas atribuídas ao

mesmo grupo caracterizem uma tempestade elétrica. Como há grande variabilidade do número de descargas de acordo com o período do dia, estações do ano, eventos meteorológicos atuantes, entre outros, não é possível utilizar técnicas de clusterização que necessitem informar o números de *clusters*. Esta limitação reduz significativamente o número de métodos viáveis para o problema proposto.

As fases iniciais do trabalho de Woolford e Braun (2006) e a presente pesquisa apresentam similaridade, isto é, ambas as pesquisas primeiramente tem o propósito de organizar as descargas atmosféricas no espaço e no tempo utilizando análise de *cluster*. Os referidos autores utilizam como algoritmo de clusterização o método *Convergent Data Sharpening*, desenvolvido pelos próprios autores (foram eles quem implementaram e atualizam o pacote *CHsharp* do R, o qual será utilizado nesta pesquisa).

Uma vantagem da utilização de tal método para caracterizar tempestades elétricas é que ele representa todos os pontos de um determinado *cluster* em um único valor, que é a moda. Essa característica é diferenciável para esta aplicação pois não se está apenas buscando fazer simples clusterizações, mas sim identificar núcleos ativos de tempestades elétricas que se conectem com núcleos de períodos seguintes para desempenhar a sequência das tempestades. O valor modal, representando o centro da tempestade, favorece o reconhecimento do percurso predominante do evento meteorológico. O método *Convergent Data Sharpening* será visto a seguir.

3.4.1 *Convergent Data Sharpening*

Convergent Data Sharpening é uma técnica baseada em uma forma convergente do método conhecido como *Data Sharpening* introduzido originalmente por Choi e Hall (1999). *Data Sharpening* consiste em reduzir o viés na estimação de funções densidades por meio de regressão local constante.

Considere os dados originais $x_1, x_2, \ldots, x_n$ com densidade desconhecida $f(x)$. Uma

estimativa para a densidade em um ponto x é dada por:

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} k_h(x_i - x) \tag{6}$$

onde k_h é uma função densidade de probabilidade simétrica, também chamada de função *kernel*, com parâmetro de escala h (largura da banda), que tem a finalidade de suavizar a curva ajustada, por isso também recebe o nome de parâmetro de suavização. Essa técnica de estimação da função densidade é chamada de KDE (do inglês *Kernel Density Estimation*).

Uma escolha comum para a função *kernel* é a Normal Padrão (SHEATHER, 2004), assim o gradiente de $\hat{f}(x)$ é dado por:

$$\nabla\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} k_h(x_i - x)(x - x_i) \tag{7}$$

e fazendo $\nabla\hat{f}(x) = 0$ (buscam-se pontos estacionários) e resolvendo para x, tem-se os dados ajustados $\dot{x}_1, \dot{x}_2, \ldots, \dot{x}_n$, onde:

$$\dot{x}_j = \frac{\sum_{i=1}^{n} k_h(x_i - x_j)x_i}{\sum_{i=1}^{n} k_h(x_i - x_j)} \tag{8}$$

e assim a densidade estimada é feita sobre os dados ajustados, dada por:

$$\hat{f}_s(x) = \frac{1}{n}\sum_{i=1}^{n} k_h(\dot{x}_i - x) \tag{9}$$

onde $\hat{f}_s(x)$ tem menor viés do que $\hat{f}(x)$ (CHOI; HALL, 1999; WOOLFORD; BRAUN, 2006).

Sendo assim, a ideia principal do método *Data Sharpening* é que cada dado original mova-se para mais perto de modos locais (pontos estacionários de $\hat{f}(x)$ encontrados igualando o seu gradiente a zero), visto que a função expressa em 6 tende a subestimar densidades nos picos e superestimar nos vales (CHOI; HALL, 1999).

Woolford e Braun (2006) propõem o seguinte teorema, juntamente com sua demonstração:

Teorema. *Para h fixo e para qualquer vetor inicial de observações x_0, o algoritmo Data*

Sharpening de Choi e Hall (1999) converge para um único vetor $\dot{x}$.

Demostração: Ver em Woolford e Braun (2006).

A ideia visual do método *Convergent Data Sharpening* é ilustrada na Figura 6, que mostra a sequência dos resultados após aplicação do método *Data Sharpening* em um exemplo com uma amostra aleatória. Ao longo das iterações, os dados (plotados no eixo $\dot{x}$) se tornam mais agrupados, tornando os picos e vales cada vez mais evidenciados. Na quinta imagem, nota-se que são encontrados três valores modais (picos), que representam os centros dos clusters.

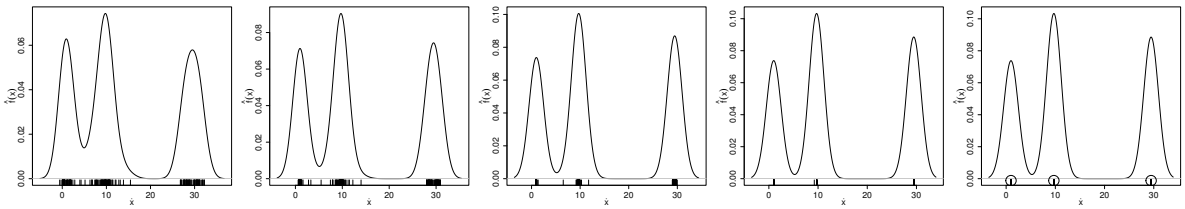


FIGURA 6: Exemplo visual de como o método *Convergent Data Sharpening* trabalha. Cinco iterações apontam três clusters encontrados em uma amostra aleatória

FONTE: A autora (2015)

Observações: Para onde os dados ajustados $\dot{x}_1, \dot{x}_2, \ldots, \dot{x}_n$ convergem, depende do suporte da função [3] *kernel* $k_h(x)$. Se o suporte é não compacto, tal com a densidade Gaussiana, os dados ajustados irão convergir para um único modo (global). Se o suporte de $k_h(x)$ é compacto, os dados ajustados podem convergir para um conjunto consistindo de mais de um ponto. Quando utilizada uma função *kernel* com suporte no intervalo $[-h;h]$ em qualquer estágio da iteração, um ponto distante de todos os demais pontos por mais que h unidades não será mais alterado. Assim, o número de modos locais detectados por este procedimento é altamente dependente do parâmetro h (WOOLFORD; BRAUN, 2006).

[3] Suporte de uma função é o menor subconjunto fechado do domínio onde a função não é nula.

A Escolha do Parâmetro de Suavização h

Conforme mencionado, o método de estimativa de densidade dado pela Equação 6 (consequentemente o método *Convergent Data Sharpening*,) é fortemente impactado pela escolha do parâmetro de suavização h. Especificamente se tratando da clusterização proposta, um valor pequeno de h resulta em uma estimativa mais fiel aos dados, gerando mais picos e vales e consequentemente mais *clusters*; um grande valor de h acarreta em uma estimativa mais suave dos dados, originando menos *clusters* (Figura 7). Assim, a escolha deste parâmetro é fundamental para o bom desempenho do sistema proposto e como essa escolha não é uma tarefa simples, h será escolhido por meio de um processo de otimização descrito posteriormente.

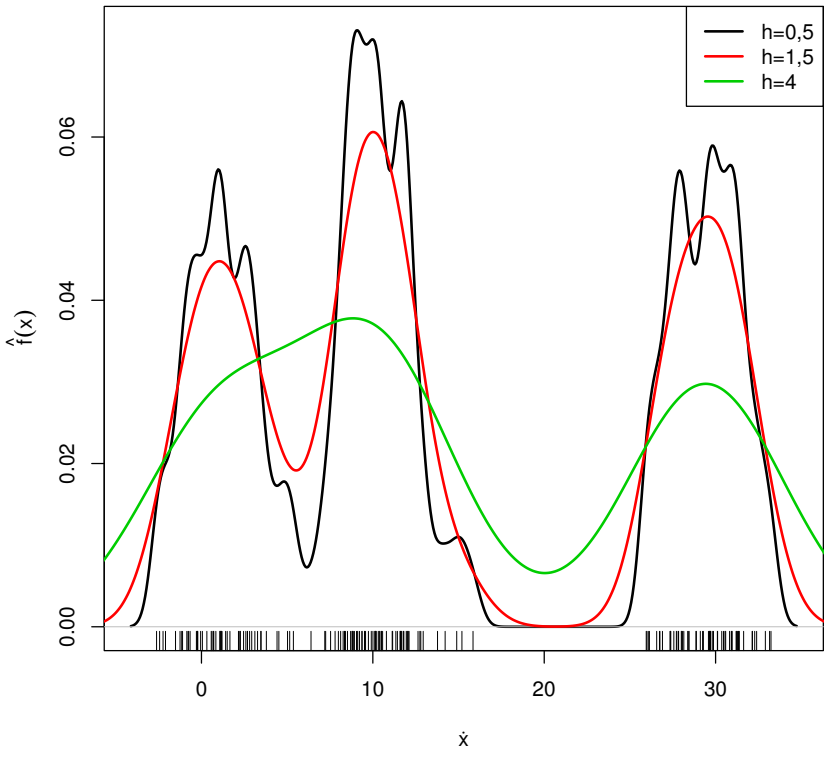


FIGURA 7: Influência do parâmetro h na estimativa da densidade
FONTE: A autora (2015)

3.5 ÍNDICES DE VALIDAÇÃO DA CLUSTERIZAÇÃO

Para avaliar de maneira objetiva e quantitativa os resultados de análise de agrupamento, são usados diferentes procedimentos a fim de se obter uma validação. Geralmente a maneira quantitativa com que se dá um procedimento de validação é alcançada por meio de um índice ou critério de validade. Tais índices/critérios podem ser de três tipos (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2001; ALBUQUERQUE, 2013):

- Externo: analisa o grau de correspondência entre a estrutura de grupos, sob avaliação pré-definida, na forma de uma solução de agrupamento esperada ou conhecida. Exemplos: Rand e Jaccard;
- Interno: analisa o grau de compatibilidade entre as estruturas de grupos sob avaliação e os dados, usando apenas os próprios dados. Exemplos: Davies-Bouldin, Dunn e Silhueta;
- Relativo: analisa várias estruturas de agrupamento do mesmo conjunto de dados, obtidas da aplicação do algoritmo com diferentes parâmetros ou ao conjunto de dados perturbados por pequenas alterações mensuráveis. Exemplos: em geral, pode ser qualquer um dos índices mencionados acima.

O propósito dos índices de validação nesta pesquisa é atestar que os *clusters* formados são adequados para a finalidade de representar núcleos de tempestades elétricas utilizando os dados de descargas atmosféricas.

O primeiro dificilmente é empregado pois o problema de clusterização é não supervisionado, isto é, não se dispõe de uma partição de referência pra validar a estrutura de grupos obtida. O terceiro não atende o propósito de avaliar um único método de clusterização com valores do parâmetro de entrada fixo (pois será otimizado conforme descrito adiante). Portanto, neste trabalho para validar a metodologia de clusterização proposta, um índice interno é utilizado. O índice Silhueta é simples e de fácil interpretação e será utilizado nesta pesquisa.

Silhueta: cada *cluster* é representado por uma silhueta (ROUSSEEUW, 1987), mostrando quais pontos se posicionam bem dentro do *cluster* e quais meramente ficam em uma posição intermediária. Para cada ponto, a seguinte informação é fornecida:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \tag{10}$$

onde a_i é a dissimilaridade média do objeto i em relação a todos os outros objetos do *cluster* A que contém i e onde b_i é a dissimilaridade média do objeto i em relação a todos os outros objetos do *cluster* mais próximo B (vizinho do objeto i). O *cluster* vizinho é uma espécie de segundo melhor *cluster* para o objeto i. Quando o *cluster* A contém apenas um objeto i, o s_i é zero.

A Equação 10 pode ser reescrita como:

$$s_i = \begin{cases} 1 - a_i/b_i, & \text{se } a_i < b_i \\ 0, & \text{se } a_i = b_i \\ b_i/a_i - 1, & \text{se } a_i > b_i \end{cases} \tag{11}$$

sendo fácil perceber que $-1 \leq s_i \leq 1$.

A largura média de silhueta de um *cluster* é a média de s_i para todos os objetos i em um *cluster*. O coeficiente de silhueta (s) é a largura média máxima da silhueta ao longo de todo k para o qual a silhueta pode ser construída. Esse coeficiente é uma medida adimensional da qualidade da estrutura de agrupamento que foi descoberta pelo algoritmo de classificação:

$$s = \max_k s_k \tag{12}$$

Rousseeuw (1987) propôs a seguinte interpretação do coeficiente s:

$s \in [0,71; 1,00]$: uma estrutura forte foi encontrada;

$s \in [0,51; 0,70]$: uma estrutura razoável foi encontrada;

$s \in [0,26; 0,50]$: a estrutura é fraca e pode ser artificial;

$s < 0,26$: nenhuma estrutura substancial foi encontrada.

4 IDENTIFICAÇÃO, MONITORAMENTO E PREVISÃO DE TEMPESTADES ELÉTRICAS

O monitoramento e previsão de eventos meteorológicos é uma tarefa complexa, pois envolve diversas incertezas relacionadas a qualidade de dados, imprecisão de métodos numéricos aplicados e, muitas vezes, do próprio comportamento desordenado dos fenômenos naturais. Betz *et al.* (2008) enaltecem que carências de medições nos dados (radar, detecção de descargas, imagens de satélites, etc), fusões e divisões entre tempestades e células com curta duração tornam o problema de monitoramento de tempestades mais complicado. Muitos trabalhos foram desenvolvidos no sentido de monitoramento e previsão de eventos meteorológicos e serão brevemente relatados a seguir.

O *Auto-Nowcast System* (ANC) (MUELLER *et al.*, 2003) é um sistema de *software* que produz, de modo geral, previsões de localização de tempestades a curto prazo (até uma hora). ANC é capaz de identificar e caracterizar camadas limite de linhas de convergência. Informações de camadas limite são usadas juntamente com características de nuvens e tempestades a fim de aumentar extrapolação com previsão de iniciação, crescimento e dissipação de tempestades. Uma rotina de lógica *fuzzy* combina campos de previsão que são baseados em observações, um modelo numérico de camada limite e sua adjunta, previsor de entrada, e algoritmos de detecção de características. Verificações estatísticas mostram que ANC é capaz de melhorar sobre extrapolação e persistência.

O algoritmo *Thunderstorm Identification, Tracking, Analysis, and Nowcasting* (TITAN) desenvolvido por Dixon e Wiener (1993) monitora e faz previsão a curto prazo de tempestades com base em dados de radar. A etapa de identificação de uma tempestade se dá em regiões contíguas que excedem limites de refletividade e volume, cujo melhor trajeto é caracterizado por meio da resolução de um problema de otimização

que visa a minimização de uma função que mede a distância percorrida e a diferença no volume. Esta minimização é fundamentada na hipótese de que em pequenos intervalos de tempo uma tempestade não se propaga longas distâncias e não altera muito sua forma. Neste processo são permitidas fusões e cisões entre tempestades. A previsão é realizada para algumas variáveis das tempestades por meio de um modelo de regressão linear onde utiliza observações anteriores que são ponderadas com pesos exponencialmente decrescentes.

Bonato (2014) utiliza rotinas de manipulação de dados brutos do TITAN para interpolar e mudar o sistema de coordenadas dos dados de radar para então comparar o desempenho de duas técnicas de previsão de células de tempestade a curtíssimo prazo: *ForTraCC* (*Forecasting and Tracking the Evolution of Cloud Clusters*) adaptado para serem utilizados dados de máxima refletividade do radar e a técnica baseada em correlação espacial, denominada técnica *XCorr*, que faz uso do método *K-Means++* para classificação de células convectivas e estratiformes. As duas técnicas foram então comparadas, por meio de gráficos de correlação entre dados observados e previsões de até uma hora geradas. A técnica *ForTraCC* apresentou melhor desempenho em relação a técnica *XCorr*.

A assimilação de dados é a técnica pela qual as observações são combinadas com a previsão numérica de tempo para aprimorar a estimativa inicial do estado da atmosfera, ou seja, melhorar a representação da realidade no instante inicial, e consequentemente um modelo capaz de prever eventos severos com maior precisão e acurácia. Inouye (2014) utiliza assimilação de dados aos modelos numéricos de previsão de tempo de mesoescala, que são inicializados a partir de condições iniciais e de contorno de um modelo global. A assimilação permite incorporar informações locais, tais como dados de radar e estações meteorológicas, para simulações de eventos de difícil previsão. Casos de estudo são apresentados a fim de averiguar se a assimilação foi vantajosa frente à rodadas do modelo sem assimilação e também ao TITAN. Como resultado, a assimilação influenciou no modelo nas três primeiras horas de modelagem, melhorando a simulação de fenômenos convectivos quando o mesmo era

detectado nas imagens de radar. Ao contrário, quando não houve a identificação de iniciação convectiva nos dados de refletividade, as simulações apresentaram desempenho similar às rodadas sem assimilação. Já o TITAN representou os núcleos com uma forma mais próxima ao observado, porém não foi capaz de acompanhar a fase do fenômeno.

Thunderstorms Radar Tracking (TRT) (HERING *et al.*, 2004) é um algoritmo que utiliza dados de três radares para detecção, monitoramento e previsão de intensos sistemas convectivos na região dos Alpes suíços. O método de identificação é baseado em um limiar de refletividade de imagens de radar que permite a detecção de células convectivas de maneira individual, dependendo do estágio do seu ciclo de vida. Para cada célula é selecionado o menor limite de refletividade que permite distingui-la a partir de células próximas, se é apresentada uma taxa de refletividade suficientemente grande. Se um limite baixo é fixado, este deve conduzir a clusterização em um único objeto de grandes áreas, incluindo células com valores de refletividade muito diferentes; se um alto valor para o limite é fixado, devem ser detectados somente núcleos de refletividade em células maduras. Por isso o algoritmo TRT propõe a utilização de três diferentes limiares para a refletividade, baseados na detecção, mínimo valor e extensão vertical. O monitoramento de células de tempestades é realizado por sucessivas imagens de radar, levando em conta sua velocidade de deslocamento. Células de períodos consecutivos são conectadas por meio da advecção da primeira de acordo com sua velocidade de deslocamento e se a área de sobreposição da advecção com a segunda célula é suficiente. Divisões e fusões de células de tempestades também são consideradas.

No TRT, até uma hora de previsão do movimento das tempestades é realizada, extrapolando o movimento de células individuais por meio da sua velocidade de deslocamento ponderada, a fim de fornecer a posição esperada. Cada célula detectada é descrita como um objeto meteorológico, e diversos atributos podem ser calculados, tais como localização, área, velocidade, taxa de crescimento, entre outros. É mencionado que extrapolação linear para métodos de previsão pode ser confiável para

algumas dezenas de minutos, dependendo da situação meteorológica. Uma vez que a extrapolação depende da velocidade, a regularidade no último passo de tempo da história dos seus atributos é determinante para o sucesso da previsão. Trajetórias regulares indicam uma previsão eficiente para algumas dezenas de minutos, enquanto que trajetórias com direções e velocidades variadas indicam previsões não confiáveis.

Desde 2003, o TRT é utilizado como ferramenta de previsão em tempo real na região alpina da Suíça. Os meteorologistas recebem de forma automatizada informações de células ativas em tempo real, as quais podem ser usadas como auxílio em tomadas de decisão em alertas. O uso de uma metodologia automatizada para detecção, rastreio e previsão de células de tempestades é uma melhoria considerável com respeito principalmente a avaliação visual subjetiva de imagens de radar. Pontos fracos do algoritmo também são relatados. Em alguns casos, tais como grandes células com fraca intensidade, o cálculo da velocidade do deslocamento da célula de tempestade se mostrou inconsistente. Os autores do referido trabalho sugerem que adicionando dados de descargas atmosféricas, o método pode melhorar no sentido de discriminar mais claramente sistemas convectivos e assim dar uma sugestão complementar da fase de vida da tempestade.

Steinacker *et al.* (2000) propõem um método para monitorar células convectivas e complexas utilizando dados de radar e de descargas nuvem-solo, de modo independente. A região de aplicação do método são os Alpes da Áustria, onde ocorrem frequentes colapsos nas estações de radar, por isso também são utilizados os dados de descargas. Além disso, os autores fundamentam-se que descargas indicam intensa convecção com suficiente eletrificação das nuvens enquanto que a refletividade medida pelo radar defronta-se com algum tipo de precipitação (convectiva assim como estratiforme) e a combinação destas duas informações permite uma distinção entre os casos. Também se considera que o ciclo de vida de uma tempestade apresenta uma típica evolução, por exemplo, descargas tendem a iniciar antes do desenvolvimento e cessam logo depois, enquanto que precipitação geralmente atinge o chão cerca de 10 a 20 minutos após as primeiras descargas, mas pode estender-se além do está-

gio maduro da tempestade. Deste modo, a avaliação combinada destes dois dados permite, além do próprio monitoramento, o reconhecimento precoce do estágio de desenvolvimento de uma tempestade. Para os dois tipos de dados são criadas grades regulares com medições em intervalos de 20 minutos, assim dois campos discretos são comparados: densidade de descargas e taxa de precipitação.

O princípio da metodologia empregada no trabalho de Steinacker *et al.* (2000) é a utilização de um filtro Gaussiano, o qual tem um parâmetro HW (*half width*) que define a largura desse filtro. Valores pequenos para este parâmetro resultam na identificação de células convectivas e valores mais altos detectam somente células complexas. Para conectar células de períodos seguidos, para cada célula individual identificada são calculados os possíveis vetores deslocamento que ela pode seguir e, segundo alguns critérios, é selecionado o melhor vetor deslocamento para cada célula. Os autores adotaram apenas fusões são realizadas, divisões não são permitidas. Três casos de estudo são apresentados no trabalho. No primeiro caso, apenas os dados de descargas são utilizados, onde dois parâmetros para definir a largura do filtro são empregados. O parâmetro com valor pequeno identificou diversas células convectivas e o parâmetro grande identificou algumas células complexas, inclusive com fusões bastante evidentes. No segundo exemplo, para um parâmetro grande do filtro, são comparados os resultados do método utilizando os dados de radar e de descargas. Apesar das trajetórias não serem idênticas, elas apresentam certa similaridade. Em alguns casos as células de tempestades são identificadas mais cedo com dados de descargas do que com os dados de radar. Houve também uma falha no rastreio de uma tempestade com os dados de descargas em que os dados de radar fecharam a lacuna formada. No último exemplo, utilizando um parâmetro pequeno no filtro, os dados de descargas mostram intensa atividade elétrica na região, porém com curtas durações, não muito organizadas. Aumentando o valor do parâmetro e utilizando as duas fontes de dados, os resultados mostram que ambos os dados rastreiam bem uma célula complexa vindo do oeste para o leste, porém em um certo momento uma trajetória desvia para o nordeste e outra para o sudeste, sugerindo que aquela posição

seria um ponto de divisão da célula.

Bonelli e Marcacci (2008) apresentam dois sistemas capazes de prever e emitir alertas de tempestades severas no norte da Itália, a fim de reduzir riscos à vida humana e prejuízos em diversos setores, tais como o elétrico e telecomunicações. Estes sistemas se baseiam na detecção, rastreio e previsão de tempestades para um curto período de tempo (meia hora), um utilizando dados de precipitação pluvial estimada por um radar e o outro com dados de descargas atmosféricas do tipo nuvem-solo. Os autores ressaltam que para a detecção da tempestade, dados com alta resolução são fundamentais. A precipitação é tomada em cada ponto de grade de uma malha de 1 km a cada 15 minutos. Já para as descargas, em cada ponto de grade de uma malha de 5 km é contabilizado o número de incidências a cada 15 minutos (a malha para as descargas é maior que a do radar porque o número de descargas que caem dentro de cada ponto de grade é muito pequeno). Uma célula de tempestade é localizada, na grade (radar ou descarga) por meio de um procedimento começando do máximo valor de um pixel (precipitação ou número de descargas associado a cada ponto de grade) em toda a grade e busca-se encontrar uma estrutura contínua de pixel em torno dele. Quando o próximo valor do pixel encontrado na estrutura cai abaixo de um limite fixado, então a célula é definida e diversos atributos podem ser calculados. Outra célula pode ser detectada da mesma forma eliminando previamente todos os pixels da primeira célula da grade. O procedimento acaba quando nenhuma estrutura de mais de 4 pixels permanece na grade.

Ainda em referência ao trabalho de Bonelli e Marcacci (2008), casos foram apresentados e mostram a boa atuação do sistema utilizando dados de radar. Também é apresentada a diferença, em um exemplo, entre os sistemas que utilizam dados de radar e dados de descargas: em alguns momentos os sistemas se mostram similares, porém em outros momentos os sistemas apresentam diferenças significativas. São justificadas as discrepâncias de direção e velocidade às incertezas na determinação do centro da célula (necessário para o cálculo dos atributos das tempestades) com pequeno número de descargas ou grandes células. Foi observado em poucos expe-

rimentos que o monitoramento de tempestades com dados de descargas tem melhor desempenho quando grandes sistemas de mesoescala estão sendo rastreados. Ainda no trabalho, é apresentado um caso de estudo onde foi detectado um elevado número de descargas positivas (25% do total) durante um tornado na mesma região de estudo, sendo que somente de 1 a 2% do total de células detectadas naquele dia apresentam este número de descargas positivas. Porém os autores ressaltam a necessidade de se analisar mais casos. Além disso, um sistema de alerta de tempestades severas foi testado na estação quente de 2007, emitindo alertas por mensagem de celular para responsáveis da Defesa Civil de 12 cidades do norte da Itália. Após poucos minutos de tempo de processamento para detectar células de tempestades, compará-las com suas precursoras e extrapolar (30 minutos à frente) sua posição usando a velocidade calculada, o sistema verifica se uma tempestade severa (determinada por um limiar de precipitação) sobrepõe uma área pré-estabelecida. Das 12 cidades averiguadas, em apenas uma os resultados são abaixo do esperado, porém esta cidade não apresenta boa cobertura do radar por pertencer a uma região montanhosa. Para concluir, os autores se declaram satisfeitos com os resultados obtidos pelos dois sistemas testados, tanto com dados de radar quanto com dados de descargas atmosféricas. Ressaltam que células de tempestades desenvolvidas localmente e com vida curta não podem ser monitoradas com precisão.

No trabalho de Betz *et al.* (2008) é proposto um método de identificação e monitoramento de células de tempestades em uma região da Alemanha utilizando dados de descargas do tipo intra-nuvens e nuvens-solo. O processo de identificação de células de tempestade é feito por meio de densidade de descargas. Na referida pesquisa, uma célula é dita suspeita quando o número de descargas por área supera um conjunto mínimo e bordas são reconhecidas quando o número absoluto de descargas fica abaixo deste mínimo. A cada 10 minutos (sem sobreposição) os dados são inspecionados e células são identificadas. Uma célula se relaciona com uma imagem anterior quando localização e velocidade esperadas estão dentro de limites selecionados. Desde que o movimento seja frequentemente linear, previsão da posição de célula para um tempo

até 2 horas é possível. Os autores apresentam casos de estudo e quatro parâmetros das tempestades com respeito as descargas são calculados: taxa de descarga, densidade de descarga, velocidade e área da célula. O primeiro exemplo serve para demonstrar as opções disponíveis para a trajetória de uma célula, onde claramente fusões e divisões poderiam ocorrer, porém não são tratadas devido ao argumento dos autores de que fusões e divisões de células geralmente levam a saltos não muito realísticos dos parâmetros das células. Também são analisados os parâmetros de uma célula especifica e constata-se, pela fraca variação dos parâmetros ao longo do tempo, que esta célula não apresentou risco de ocasionar evento severo. Outro exemplo é apresentado para um dia com ocorrência de evento severo, com ocorrência de granizo, ventos fortes e intensa precipitação. São estudados os parâmetros mencionados e constatado que uma célula teve rápido aumento da taxa de descarga no tempo próximo a danos terrestres significantes. O último exemplo mostra a relação entre descargas e radar. É observado que refletividade acima de 30 dBZ[1] está associada com descargas, porém algumas descargas ocorrem fora de áreas de alta refletividade, e vice-versa. É ressaltada a importância de um estudo mais aprofundado nesta linha de pesquisa. Os autores mencionam que, quanto ao monitoramento de células com dados de descargas, uma simples extrapolação baseada em três ou quatro intervalos de tempo consecutivos de cinco ou dez minutos produzem resultados eficientes, especialmente quando células apresentam vida longa.

Strauss, Rosa e Stephany (2013) propõem uma nova técnica baseada em clusterização espacial-temporal com janela deslizante no tempo e KDE para identificar e rastrear células eletricamente ativas de descargas nuvem-solo. O processo de clusterização, realizado pelo algoritmo DENCLUE (*Density-based Clustering*), identifica grupos de descargas que correspondem a diferentes células de atividade elétrica. É realizada por meio da posição das descargas em uma janela de largura fixa que desliza em um certo passo de tempo. Os parâmetros da janela utilizados no trabalho são:

[1] dBZ significa decibéis de Z, onde Z é o fator de refletividade do radar meteorológico, Z é diretamente relacionado com os diâmetros de gotas de chuvas presentes na atmosfera. Maiores detalhes podem ser encontrados em Rinehart (2004).

10 minutos de largura e passo de 5 minutos (tempos compatíveis com os dados de radar que são interpolados em intervalos de 10 minutos). Sempre que necessário, divisões e uniões são realizadas. O método KDE é então aplicado a cada *cluster* formado separadamente, resultando em um conjunto de campos de densidade de ocorrências de descargas nuvem-solo.

O Estado de São Paulo é a região escolhida para aplicação da metodologia proposta no trabalho mencionado, onde o sistema de detecção de descargas tem eficiência de 90%. Dois eventos convectivos ocorridos em janeiro de 2010 foram selecionados para estudo. São ilustradas as células com atividade elétrica identificadas apenas utilizando KDE e também pela técnica proposta pelos autores, ou seja, aplicação da clusterização espacial/temporal das descargas com janela deslizante e passo de tempo fixo e em seguida a aplicação da KDE a cada célula identificada. Os resultados mostram que a combinação de clusterização e KDE permite a identificação e visualização mais precisa de células com atividade elétrica, uma vez que apenas a técnica KDE tende a ampliar as células e também gerar células consideradas "falsas positivas", visto que não há precipitação convectiva associada a elas, utilizando imagens de radar para isto. Uma análise de correlação é feita entre imagens de radar e células com atividade elétrica identificadas somente por KDE e também pela combinação de clusterização e KDE. O resultado mostra que a combinação de métodos proposta resulta em melhor correlação com imagens obtidas pelo radar do que somente utilizando a técnica KDE, visto que é sugerida uma forte correlação entre descargas atmosféricas e ocorrências de células convectivas.

Woolford e Braun (2006) utilizam clusterização de descargas atmosféricas para rastrear os centros de tempestades elétricas no espaço e no tempo. Os autores mencionam que “o processo de clusterização é um modelo natural que considera descargas ocorrendo em grupos espaciais/temporais como sistema de tempestades movendo-se ao longo de uma área”. A região de aplicação é a província de Ontário no Canadá, onde de 35 a 40% dos incêndios florestais são causados por descargas atmosféricas e a meta da pesquisa é desenvolver uma metodologia para mapear a probabilidade

de ocorrência de incêndios florestais usando informações de descargas.

No trabalho é apresentado um exemplo com mais de 1 milhão de descargas referentes ao ano de 1994 em que as descargas foram clusterizadas pelo método *Convergent Data Sharpening* no espaço (com parâmetro h de 1° ou equivalentemente a 100 quilômetros) e no tempo (com parâmetro h de 1 dia). Como resultado, pôde-se classificar incêndios florestais causados por descargas se está próximo a um centro de tempestade elétrica formada pela clusterização.

4.1 APLICAÇÃO: IDENTIFICAÇÃO E MONITORAMENTO DE TEMPESTADES ELÉTRICAS

A detecção das tempestades elétricas é realizada por meio da clusterização da posição espacial de descargas atmosféricas pelo método *Convergent Data Sharpening*, que no *software* R está implementado dentro da biblioteca *CHsharp* com o uso da função *sharp2d* (R Core Team, 2012). A aplicação emprega janela móvel de uma hora com passo de dez minutos, ou seja, um conjunto de descargas é selecionado e clusterizado no período de uma hora, e após a clusterização a nova janela temporal considerada é a anterior com dez minutos retirados no início e dez minutos acrescidos no final e novamente um novo conjunto de descargas é selecionado e clusterizado, e assim por diante (Figura 8).

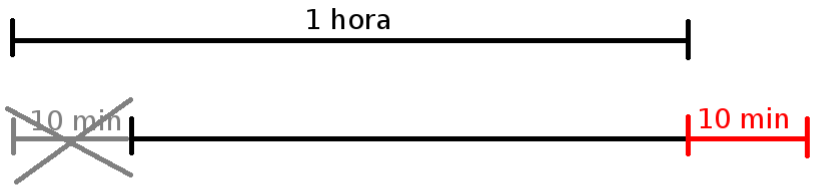


FIGURA 8: Janela móvel de uma hora, onde a cada passo, dez minutos são retirados no início e dez minutos são acrescidos no final da janela

FONTE: A autora (2015)

A clusterização aplicada conforme descrita origina retratos temporais, onde uma mesma descarga encontra-se em seis *clusters* distintos. Estes retratos temporais devem ser conectados (por meio de alguma medida de similaridade) uns aos outros

para caracterizar o percurso das tempestades. Na pesquisa, *clusters* de janelas consecutivas são unidos se a velocidade de deslocamento de um núcleo para outro não ultrapassa um limite máximo, estipulado em 50 km/10min ou equivalentemente 300 km/h. Em outras palavras, em 10 minutos é permitido um deslocamento dos centros dos *clusters* de janelas consecutivas de até 50 km. Isso não significa que o sistema trabalha com velocidade de deslocamento de tempestades da ordem de 300 km/h (o que é superior à encontrada na literatura para a região (BENETI, 2012)), mas é um procedimento utilizado para que as tempestades tenham uma sequência temporal mais contínua e para tratar da instabilidade inicial e final na formação dos *clusters*. Foram testadas variações de 10 a 100 km/10min para este parâmetro, porém valores de 10 a 40 km/10min conduzem a divisões inadequadas de uma tempestade e valores acima de 50 km/10min, apesar de não alterar significativamente as conexões criadas nesta aplicação, pode agregar tempestades distintas. Feito isso é concluído o procedimento de identificação e rastreamento das tempestades elétricas. Por convenção, adotou-se o critério que uma tempestade tem memória máxima de 3 horas, visto que é um período de tempo suficiente para uma tempestade passar do estágio inicial de formação para o estágio de maturação. Assim, se uma tempestade durar mais que 3 horas, o passado mais antigo é descartado, dando lugar ao histórico mais recente da tempestade.

Neste processo de união de *clusters*, três situações distintas podem ocorrer: fusão, cisão e formação comum de tempestades.

- **Fusão** (FIGURA 9): duas tempestades estão ativas no tempo $t-1$. No tempo t elas se fundem, porém o histórico que a nova tempestade irá herdar será o histórico da tempestade cujo centro é mais próximo do seu centro. A outra tempestade acaba no tempo $t-1$.

- **Cisão** (FIGURA 10): uma tempestade está se desenvolvendo. Porém em um certo tempo t ela se desmembra em duas. Todo histórico da tempestade acom-

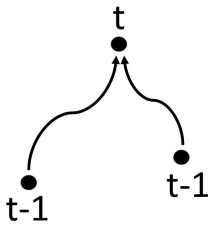


FIGURA 9: Fusão de duas tempestades
FONTE: A autora (2015)

panha àquela cujo centro esteja mais próximo do centro da tempestade no tempo $t-1$. A outra tempestade começará um novo histórico no tempo t.

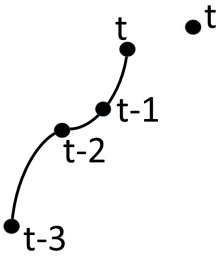


FIGURA 10: Cisão de duas tempestades
FONTE: A autora (2015)

- **Tempestades comuns** (FIGURA 11): uma tempestade começa e acaba sem sofrer fusão ou cisão.

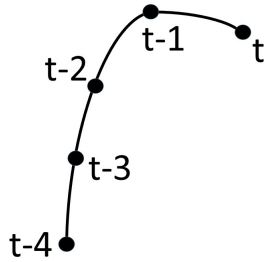


FIGURA 11: Formação comum de uma tempestade
FONTE: A autora (2015)

Duas observações importantes sobre a forma de aplicação da clusterização empregada nesta pesquisa são:

1. A clusterização de um conjunto de dados pelo método *Convergent Data Sharpening* foi realizada de modo que os *clusters* formados permanecessem inalterados por cinco iterações consecutivas, ou seja, o número de repetições do método *Data Sharpening* não é fixo e varia de acordo com a amostra de pontos e sua pertinência nos *clusters*. Se, em cinco iterações seguidas, os *clusters* formados não sofrem alterações, então a clusterização termina. Este critério de parada para o método *Convergent Data Sharpening* é bastante simples e induz que dificilmente haverá modificações significativas caso o algoritmo continuasse. Também ressalta-se a vantagem da utilização do método *Convergent Data Sharpening* pelo fato de não precisar informar previamente o número de *clusters*, fundamental nesta aplicação, já que o número de dados das janelas móveis varia muito de acordo com estação do ano, período do dia, etc;

2. As clusterizações de janelas móveis consecutivas não são independentes, isto é, a estrutura de agrupamento de dados de uma janela é aproveitada na janela seguinte, visto que muitos dados antigos permanecem no conjunto a ser clusterizado no novo período, conforme já relatado. Este aproveitamento ocorre da seguinte maneira: dados antigos (de 10 minutos iniciais da clusterização anterior) são removidos do conjunto amostral; cada dado da clusterização anterior que permanece no conjunto amostral é substituído pelo valor médio dos pontos do *cluster* ao qual pertence; novos dados (dos últimos 10 minutos) são introduzidos na amostra a ser clusterizada. Esse artifício de rearranjo dos dados antes da clusterização seguinte, além de acelerar o processo de convergência do método, favorece uma especie de continuidade da clusterização anterior, permitindo uma maior capacidade de conexão entre *clusters*, promovendo a sequência das tempestades e assim um acompanhamento mais realista dos eventos meteorológicos e consequentemente das variáveis analisadas. Desta forma, novos *clusters* gerados retratam o deslocamento temporal de *clusters* passados. Além disso, a própria característica do método *Convergent Data Sharpening*, que é a aderência dos pontos a modos locais, propicia a identificação de núcleos de

tempestades que seguem o movimento dominante de eventos meteorológicos.

Contudo, para aplicação da metodologia proposta, inicialmente deve-se escolher o parâmetro h do método de clusterização *Convergent Data Sharpening*. Este é o único parâmetro do sistema proposto e será determinado por meio da resolução de um problema de otimização (descrito posteriormente) que visa conciliar bom monitoramento e previsão das tempestades elétricas.

4.1.1 Atributos Analisados das Tempestades Elétricas

Após identificadas, diversos atributos das tempestades elétricas podem ser determinados ao longo do seu passado. Os atributos das tempestades calculados nesta pesquisa são:

- **Posição central**: corresponde a localização (longitude $\times$ latitude) do centro da tempestade elétrica, calculada a partir do valor médio das posições das descargas que compõem cada tempestade;
- **Número de descargas**: quantidade média de descargas em uma hora (descargas/hora) pertencente a cada tempestade;
- **Pico de corrente**: média do valor absoluto do pico de corrente das descargas (kA) que compõem cada tempestade;
- **Distribuição espacial das descargas**: representação no espaço de onde as descargas de cada tempestade se localizam (graus^2).

Para retratar a distribuição espacial das descargas dentro de uma tempestade é utilizada a estimativa de distribuição dos dados por uma função binormal, e devido as suas características produz uma elipse (chamada de elipse de incerteza) que indica uma região confiável, a um certo nível, onde as descargas se encontram no espaço. A área desta elipse simboliza a região de abrangência espacial da tempestade. Esta elipse, e portanto a distribuição espacial de uma tempestade, é representada por meio

de uma única matriz de covariância 2x2, a qual é capaz de fornecer todos os atributos da elipse (eixos maior e menor, centro e ângulo) que geram a área desta elipse. Esta matriz de covariância é calculada pela função *Mclust()* da biblioteca *mclust* do R, que é um método de clusterização por modelos, brevemente descrito na Seção 3.3.1, baseado na ideia de ajustar uma função densidade de probabilidade sobre os dados por meio de misturas de funções gaussianas finitas (FRALEY; RAFTERY, 2007; FRALEY *et al.*, 2012), porém na aplicação proposta utilizou-se apenas uma função gaussiana a fim de representar adequadamente a disposição das descargas dentro de um único *cluster* (tempestade) já estruturado pelo *Convergent Data Sharpening*.

Uma observação importante a respeito ao pico de corrente é que, em toda a pesquisa, tanto na etapa de monitoramento quanto na etapa de previsão, optou-se por trabalhar com o logaritmo desta variável, pelo motivo de que é sabido que o pico de corrente das descargas segue a distribuição lognormal (BERGER; ANDERSON; KRÖNINGER, 1975) e portanto o logaritmo desta variável é aproximadamente normal. Este fato favorece alguns cálculos e torna o problema mais simples. Entretanto, nas figuras que seguem, a própria média do valor absoluto do pico de corrente das descargas será apresentada, pelo habitual uso desta variável no cotidiano e na literatura.

Todos os atributos são calculados a cada passo de tempo na etapa de clusterização e o acompanhamento destes permite monitorá-los ao longo do tempo. A previsão, para um determinado período de tempo à frente (até uma hora), indicará o comportamento do evento meteorológico estudado que é a tempestade elétrica.

4.2 APLICAÇÃO: PREVISÃO DE TEMPESTADES ELÉTRICAS

Um método amplamente utilizado na literatura para realizar previsão de variáveis meteorológicas (DIXON; WIENER, 1993; HERING *et al.*, 2004; BONELLI; MARCACCI, 2008) é a extrapolação de dados, que consiste basicamente em aproximar um valor desconhecido fora de um intervalo de pontos conhecidos. Uma curva (função) é ajustada a um conjunto de observações (uma série histórica por exemplo) e

esse padrão é estendido para o futuro. Sabe-se que o resultado da extrapolação de dados é mais confiável quando o horizonte da previsão é menor e quando a série de dados históricos apresenta comportamento mais regular (uniforme). Para uma boa previsão, informações essenciais sobre o futuro da série de dados estão contidas na série histórica e pressupõem-se que tendências passadas se estenderão para o futuro. Se o futuro apresentar novos efeitos não observados no passado, eventualmente a extrapolação será inadequada.

Portanto, para extensão de um padrão para o futuro (extrapolação), antes é preciso determinar uma curva ou função que se adeque satisfatoriamente bem ao conjunto de dados já conhecidos.

4.2.1 Ajuste de Curvas

O ajuste de curvas é uma forma de representação de um conjunto de dados por meio de uma equação matemática. A função ajustada não necessariamente precisa fornecer o valor exato em cada ponto, e sim representar adequadamente o conjunto de dados como um todo, representando a tendência geral dos dados. “Em geral, qualquer medição experimental apresenta erros ou incerteza inerentes, e a procura por uma curva que passe por todos os pontos medidos não traz consigo qualquer benefício” (CHAPRA; CANALE, 2008).

O ajuste de curva mais simples é o ajuste ou regressão linear, que consiste em aproximar os pontos por uma reta da forma $y = a_0 + a_1 x$. Se o conjunto de dados é constituído por apenas dois pontos, a_0 e a_1 são determinados de modo que a equação passa pelos dois pontos. Se o conjunto de dados é formado por mais de dois pontos, é evidente que uma reta não passará por todos os pontos (ao menos que os pontos sejam colineares), e assim os coeficientes a_0 e a_1 são determinados a fim de que a reta ajustada forneça o melhor ajuste como um todo.

A maneira de medir quão bem uma função pode representar um conjunto de dados é pelo erro ou resíduo, que consiste na diferença entre cada ponto do conjunto

de dados e o valor da função aproximada. Quando é minimizada a soma dos quadrados dos resíduos entre o valor observado e o valor calculado pelo ajuste, o método recebe o nome de regressão por mínimos quadrados (GILAT; SUBRAMANIAM, 2008; CHAPRA; CANALE, 2008).

Ajuste de Curvas com a Linearização de Equações Não Lineares

A maioria das aplicações práticas na ciência e na engenharia confirmam que a relação entre as grandezas envolvidas não é linear. Assim, o uso de funções não lineares levam a ajustes muito melhores dos dados experimentais do que o uso de funções lineares, na maioria de problemas práticos. Porém, funções não lineares podem ser escritas de tal forma que possibilite a determinação dos coeficientes que levem ao melhor ajuste por regressão linear por mínimos quadrados.

Para utilizar a regressão linear, uma equação não linear deve ser transformada tal que a nova equação seja linear com termos incluindo as variáveis originais. Por exemplo, a função exponencial $y = be^{mx}$ pode ser linearizada aplicando-se o logaritmo natural (ln) em ambos os lados da equação:

$$ln(y) = ln(be^{mx}) = ln(b) + mx \tag{13}$$

Essa equação é linear com $ln(y)$ em termos de x. A equação está na forma $Y = a_0 + a_1X$, em que $Y = ln(y)$, $X = x$, $a_0 = ln(b)$ e $a_1 = m$:

$$\underbrace{ln(y)}_{Y} = \underbrace{ln(b)}_{a_0} + \underbrace{m}_{a_1}\underbrace{x}_{X} \tag{14}$$

Dessa maneira, uma regressão linear por mínimos quadrados pode ser utilizada para fazer com que uma equação na forma $y = be^{mx}$ se ajuste a um conjunto de pontos (x_i, y_i). Após calculados a_0 e a_1, os coeficientes b e m na equação exponencial são calculados da forma: $m = a_1$ e $b = e^{a_0}$.

Similarmente diversas outras equações não lineares podem ser linearizadas, onde

algumas destas equações estão listadas na Tabela 2.

TABELA 2: Linearização de equações não lineares

Equação não linear	**Forma linear**	**Relação com $Y = a_0 + a_1 x$**	**Valores para a regressão linear por mínimos quadrados**	**Gráficos onde os dados medidos parecem se ajustar a uma linha reta**
$y = bx^m$	$ln(y) = mln(x) + ln(b)$	$Y = ln(y)$, $X = ln(x)$, $a_1 = m$, $a_0 = ln(b)$	$ln(x_i)$ e $ln(y_i)$	Gráfico y vs. x em eixos x e y logarítmicos. Gráfico $ln(y)$ vs. $ln(x)$ em eixos x e y lineares
$y = be^{mx}$	$ln(y) = mx + ln(b)$	$Y = ln(y)$, $X = x$, $a_1 = m$, $a_0 = ln(b)$	x_i e $ln(y_i)$	Gráfico y vs. x em eixos x linear e y logarítmico. Gráfico $ln(y)$ vs. x em eixos x e y lineares
$y = b10^{mx}$	$log(y) = mx + log(b)$	$Y = log(y)$, $X = x$, $a_1 = m$, $a_0 = log(b)$	x_i e $log(y_i)$	Gráfico y vs. x em eixos x linear e y logarítmico. Gráfico $log(y)$ vs. x em eixos x e y lineares
$y = \frac{1}{mx+b}$	$\frac{1}{y} = mx + b$	$Y = \frac{1}{y}$, $X = x$, $a_1 = m$, $a_0 = b$	x_i e $\frac{1}{y_i}$	Gráfico $\frac{1}{y}$ vs. x em eixos x e y lineares
$y = \frac{mx}{b+x}$	$\frac{1}{y} = \frac{b}{m}\frac{1}{x} + \frac{1}{m}$	$Y = \frac{1}{y}$, $X = \frac{1}{x}$, $a_1 = \frac{b}{m}$, $a_0 = \frac{1}{m}$	$\frac{1}{x_i}$ e $\frac{1}{y_i}$	Gráfico $\frac{1}{y}$ vs. $\frac{1}{x}$ em eixos x e y lineares

FONTE: Gilat e Subramaniam (2008)

Escolha da Função Apropriada

Como visto, diversas funções podem ser utilizadas para ajustar um conjunto de dados, porém a escolha da função ajuste depende essencialmente dos dados, que podem ser contínuos, suaves, periódicos, entre outros. Uma análise exploratória nos dados históricos permite uma melhor compreensão de tendências e assim um modelo mais apropriado pode ser escolhido.

Uma forma de verificar a relação entre as grandezas é por meio de um gráfico de pontos do conjunto de dados, chamado gráfico de dispersão. O traçado de um gráfico

de dispersão com eixos lineares indica se esta relação é linear ou não: se o traçado se parecer com uma linha reta, então a relação entre as grandezas é linear; se o traçado se parecer com uma curva, então a relação entre as grandezas é não linear.

É possível prever, até certo ponto, se uma função não linear é adequada para representar um conjunto de dados, com o traçado dos pontos medidos de uma maneira específica, analisando se esses pontos indicam a formação de uma reta. Para as funções retratadas na Tabela 2, essa informação é apresentada na última coluna da tabela. Para todas as funções, isso pode ser feito com o traçado dos valores transformados do conjunto de dados em gráficos com eixos lineares. Adicionalmente, para as três primeiras funções da tabela pode-se ainda usar diferentes combinações de eixos lineares e logarítmicos (GILAT; SUBRAMANIAM, 2008).

A extrapolação de dados no tempo será a técnica utilizada para previsão à curto prazo de alguns atributos das tempestades elétricas. Conforme mencionado, o conhecimento prévio dos dados utilizados é a chave para uma extrapolação adequada, por isso testes empíricos e análise visual foram realizados para as variáveis que serão extrapoladas: posição central, número médio de descargas, pico de corrente médio das descargas e distribuição espacial das descargas das tempestades elétricas.

Para as variáveis relacionadas a posição central da tempestade, isto é, a latitude e a longitude, foram ajustadas as seguintes funções extrapoladoras, respectivamente:

$$f(t) = e^{a_1 t + b_1} \tag{15}$$

e

$$f(t) = e^{a_2 t + b_2} \tag{16}$$

onde t representa a variável tempo, a_1 e b_1 são coeficientes obtidos por meio da linearização da equação não linear conforme apresentado anteriormente, ou matematicamente, $ln(lat_1, \ldots, lat_n) \sim (t_1, \ldots, t_n)$ em que $lat_1, \ldots, lat_n$ são dados passados e já conhecidos. De forma análoga, a_2 e b_2 são obtidos de $ln(lon_1, \ldots, lon_n) \sim (t_1, \ldots, t_n)$. Latitude e longitude são extrapoladas independentes umas das outras.

Para as demais variáveis, a função extrapoladora utilizada é:

$$f(t) = e^{aln(t)+b} \tag{17}$$

onde novamente t é o tempo, a e b são coeficientes provenientes da linearização da equação não linear conforme apresentado anteriormente. No *software* R, o ajuste linear utilizando o método dos mínimos quadrados é feito pela função *lm()* do pacote *stats* (R Core Team, 2012).

Vale ressaltar que os coeficientes do ajuste linear não são previamente calculados e então aplicados para projetar as variáveis, mas sim computados em tempo de processamento (10 minutos) para cada tempestade, utilizando seu histórico recente e construindo um modelo no tempo, para então projetá-la uma hora à frente.

Outra questão importante a se considerar no ajuste de curvas é a quantidade de pontos conhecidos que serão utilizados na aproximação. Como o foco da aplicação do ajuste de curvas nesta pesquisa é a descoberta de uma função que gere uma boa previsão futura, poucas observações podem gerar uma extrapolação pobre, enquanto que muitas observações podem ocasionar extrapolação inadequada devido a irrelevância da utilização de dados antigos. Testes numéricos comprovaram a boa previsibilidade das variáveis analisadas utilizando quatro observações passadas, isto é, nos tempos $t = -30$, $t = -20$, $t = -10$ e $t = 0$ (totalizando 30 minutos de histórico) e almeja-se prever as variáveis das tempestades elétricas em $t = +60$.

4.3 AJUSTE DO PARÂMETRO DO SISTEMA

Conforme apresentado, o sistema de detecção, monitoramento e previsão de tempestades elétricas é dependente do parâmetro de suavização h do método de clusterização. A escolha deste parâmetro deu-se por meio da análise da eficiência de monitoramento e previsão do sistema, baseado em otimização multicritério. Porém, para tal análise de eficiência do sistema, primeiramente há a necessidade de representar as tempestades no espaço (em uma grade regular), para assim ser possível

fazer a comparação entre previsão e observação.

4.3.1 Representação das Tempestades Elétricas no Espaço

Para representar as tempestades no espaço e assim poder comparar valores previstos com observados, gerou-se uma malha regular de 10x10 km sobre a região da LT 765 kV (Figura 4) e optou-se por representar cada tempestade por uma distribuição de probabilidade do tipo normal bivariada.

Por meio desta representação bivariada, pôde-se integrar numericamente no espaço esta função de probabilidade e que cujo resultado multiplicado pela previsão do número de descargas originou o número esperado de descargas por quadrícula dado por aquela tempestade. Fazendo isto para todas as tempestades ativas em um determinado momento e somando estes números esperados, têm-se o número esperado de descargas por quadrículas dado por todas as tempestades elétricas do momento.

Também é possível estimar o valor médio do pico de corrente das descargas por quadrícula por meio do número esperado de descargas (por tempestade e por quadrícula) multiplicado pela previsão da média do valor absoluto do pico de corrente por tempestade. Somando as quantidades obtidas por cada tempestade, têm-se a média esperada do valor absoluto do pico de corrente das descargas por quadrícula influenciada por todas as tempestades ativas de um certo momento.

Finalmente estima-se a probabilidade de ocorrência de descargas por quadrícula por meio de uma distribuição binomial, calculando o complementar da probabilidade de insucesso (0 descarga por quadrícula), com parâmetros: probabilidade de sucesso dada pela integral sobre a curva normal bivariada para cada quadrícula e número esperado de descargas por quadrícula. Fazendo isto para todas as tempestades ativas em um determinado momento e multiplicando os valores, obtêm-se a probabilidade de ocorrência de descargas para cada quadrícula da grade espacial.

Estes atributos das tempestades elétricas calculados em cada ponto de grade (quadrícula) serão utilizados para comparar valores previstos e observados e assim

estimar a qualidade da previsão. Para isto, um problema de otimização multicritério é apresentado.

4.3.2 Problema de Otimização

A otimização multicritério, que definirá a escolha de um bom valor para o parâmetro h do método de clusterização, é feita da seguinte forma: para h_i, calculam-se os índices $I_n(h_i)$, $I_{PC}(h_i)$, $I_p(h_i)$ e $I_d(h_i)$ (descritos a seguir). Com estes índices, computa-se o valor da seguinte função:

$$F(I_n(h_i), I_{PC}(h_i), I_p(h_i), I_d(h_i)) = \sqrt{p_1(I_n(h_i))^2 + p_2(I_{PC}(h_i))^2 + p_3(I_p(h_i))^2 + p_4(I_d(h_i))^2} \tag{18}$$

onde p_1, p_2, p_3 e p_4 são pesos previamente definidos.

Faz-se i variar de modo que $0,1 \leq h_i \leq 1$ discriminado de $0,1$ em $0,1$ e repete-se o processo descrito anteriormente. Por fim, escolhe-se o valor de h_i que resulte o menor valor para a função F.

Os índices utilizados na função objetivo (18) são:

- I_n:

$$I_n = 1 - \frac{EMQ_{n_{ref}}}{EMQ_n} \tag{19}$$

  Índice baseado no número esperado de descargas por quadrícula, onde EMQ_n é o erro quadrático médio entre os valores previstos e observados do número de descargas por quadrícula; $EMQ_{n_{ref}} = 0,081$ é o erro que o sistema comete acertando exatamente a previsão do número de descargas das tempestades;

- I_{PC}:

$$I_{PC} = 1 - \frac{EMQ_{PC_{ref}}}{EMQ_{PC}} \tag{20}$$

  Índice baseado na média do logaritmo do valor absoluto do pico de corrente das descargas incidentes por quadrícula, em que EMQ_{PC} é o erro quadrático médio entre os valores previstos e observados desta variável por quadrícula;

$EMQ_{PC_{ref}} = 0,272$ é o erro que o sistema comete acertando exatamente a previsão do pico de corrente das descargas das tempestades;

- I_p:

$$I_p = 1 - \frac{AUC_p}{AUC_{p_{ref}}} \tag{21}$$

Índice baseado na probabilidade de ocorrência de descargas nas quadrículas, onde AUC_p é a área sob a curva ROC (JOLLIFFE; STEPHENSON, 2003) que compara o resultado da previsão (probabilidade de ocorrer descargas na quadrícula) com as verdadeiras ocorrências (ocorreu (1) ou não (0) descargas na quadrícula); $AUC_{p_{ref}} = 0,992$ é a maior área sob a curva ROC assumindo que a previsão por tempestade foi exata;

- I_d:

$$\frac{duração_{ref}}{duração} \tag{22}$$

Índice baseado na duração das tempestades, onde *duração* é o tempo médio de existência da tempestade; $duração_{ref} = 18$ representa a menor duração de uma única tempestade que abrange todas as descargas incidentes (18 momentos (máximo possível) de 10 minutos = 3 horas).

Os índices de referências de EMQ não são nulos e $AUC_{p_{ref}}$ não é igual a 1 pelo fato de que, mesmo que a previsão para as tempestades seja exatamente o real ocorrido, se têm o erro da representação elíptica das tempestades e a comparação entre valores previstos e observados é feita com base na grade gerada, isto é, por quadrícula, e não por tempestade.

Almejam-se I_n, I_{PC}, I_p e I_d pequenos, por isso a minimização da função objetivo. Os três primeiros índices estão relacionados com a parte de previsão do sistema, já o último índice é referente ao monitoramento das tempestades. Os pesos p_1, p_2 e p_3 foram ajustados com o valor $\frac{2}{9}$ (totalizando peso de $\frac{2}{3}$) e p_4 com o valor $\frac{1}{3}$, significando uma maior importância na etapa de previsão do que na etapa de monitoramento do sistema.

Como esse processo de otimização exige bastante esforço computacional (são realizadas clusterizações, conexões entre *clusters*, previsões, acúmulo do histórico das tempestades e comparação das previsões com os valores observados), um período amostral de vinte dias foi escolhido para a otimização (cinco dias seguidos para cada estação do ano de forma a representar a variabilidade climática ao longo de um ano: 05/01/2012 a 09/01/2012, 23/05/2012 a 27/05/2012, 27/07/2012 a 31/07/2012 e 23/11/2012 a 27/11/2012). O resultado da otimização é ilustrado na Figura 12, onde são retratados os índices utilizados no processo de otimização, juntamente com a função objetivo do problema quando variado o parâmetro h por $0,1$. O parâmetro bem estimado para o sistema foi $h = 0,2$ decorrente de $I_n = 0,667$ (proveniente de $EMQ_n = 0,243$), $I_{PC} = 0,432$ (proveniente de $EMQ_{PC} = 0,479$), $I_p = 0,034$ (proveniente de $AUC_p = 0,958$) e $I_d = 0,233$ (proveniente de *duração* $= 77,279$), resultando em um valor de $0,398$ para a função objetivo.

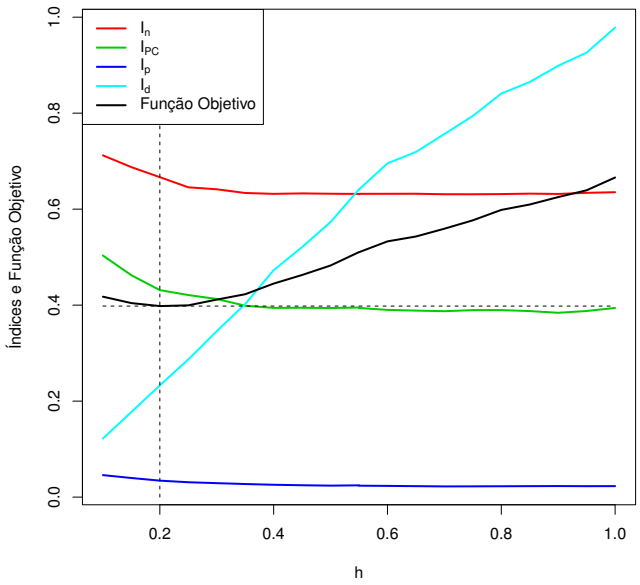


FIGURA 12: Valores dos índices e função objetivo do problema, cuja melhor solução foi $h = 0,2$

FONTE: A autora (2015)

Para os vinte dias selecionados no processo de otimização e utilizando o parâmetro $h = 0,2$, calculou-se o índice silhueta a fim de averiguar se os *clusters* encontrados

na representação dos núcleos de tempestades elétricas são satisfatórios. Conforme apresentado na Seção 3.5, o índice silhueta varia de -1 a 1, sendo que valores próximos de -1 indicam *clusters* mal estruturados e valores próximos de 1 sugerem boa representação dos dados em seus respectivos *clusters*. A Figura 13 mostra a função de distribuição acumulada empírica das médias do índice silhueta para o período de vinte dias escolhidos. Segundo a classificação de Rousseeuw (1987), 10,9% dos *clusters* não obtiveram estrutura substancial, 34,6% dos *clusters* são estruturalmente fracos, 36,2% dos *clusters* são razoavelmente estruturados e 18,3% dos *clusters* são estruturados fortemente. Os casos não muito bons podem ser justificados pela presença de ruídos, já que descargas nuvem-solo são eventos de natureza caótica.

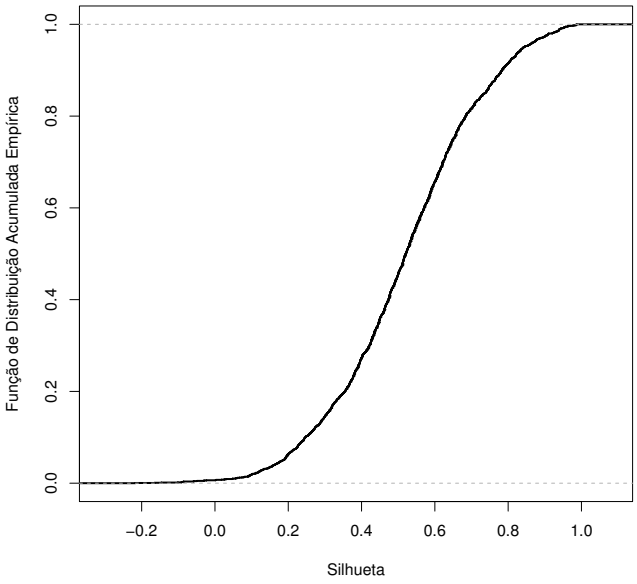


FIGURA 13: Função de distribuição acumulada empírica das médias do índice silhueta
FONTE: A autora (2015)

Validação do Sistema

Para validar o resultado do sistema proposto utilizando o parâmetro encontrado pela otimização multicritério, um novo período de 20 dias totalmente independente do teste anterior é avaliado com $h = 0,2$ a fim de verificar se a amostra utilizada na otimização é representativa. O novo período é: 05/01/2014 a 09/01/2014, 18/04/2014

a 22/04/2014, 25/08/2014 a 29/08/2014 e 09/12/2014 a 13/12/2014. Os índices resultantes são: $I_n = 0,905$ (proveniente de $EMQ_n = 0,852$), $I_{PC} = 0,486$ (proveniente de $EMQ_{PC} = 0,529$), $I_p = 0,031$ (proveniente de $AUC_p = 0,961$) e $I_d = 0,175$ (proveniente de *duração* $= 102,856$), resultando em um valor de $0,495$ para a função objetivo. Os dois primeiros índices são um pouco maiores no período de validação do que no de calibração, o terceiro índice é praticamente igual nos dois períodos e o quarto índice é menor no período de validação. A função objetivo do período de validação teve um aumento de 24,4% em relação ao período de calibração, o que pode ser considerado esperado pelo fato de que na validação a amostra é totalmente inexplorada. Assim, pode-se concluir que o período amostral utilizado para calibração do sistema é representativo.

5 RESULTADOS E DISCUSSÕES

Encontrado um bom parâmetro para o sistema ($h = 0,2$), é possível aplicá-lo para detectar, monitorar e prever tempestades elétricas de modo pré-operacional na região piloto, podendo uma tempestade elétrica ser acompanhada, inspecionando diversas características ao longo do seu desenvolvimento e do seu futuro.

O sistema proposto produz características tanto do passado da tempestade quanto do seu futuro (uma hora à frente), sendo possível acompanhar visualmente algumas destas informações. Como já mencionado, uma tempestade tem histórico máximo de 3 horas no sistema apresentado. Para as figuras apresentadas a seguir, o passo de 10 minutos na consideração da janela móvel da clusterização foi modificado para 1 minuto com o intuito de discretizar melhor o tempo e originar traçados mais contínuos. Porém esta modificação não afeta em nada o que já foi realizado, uma vez que o passo de 1 minuto não foi utilizado na etapa de ajuste do parâmetro devido ao esforço computacional.

Para ilustrar o resultado do sistema proposto, dois exemplos reais serão apresentados. O primeiro caso de estudo é referente a um período em que foi registrada uma falha no sistema elétrico piloto e o segundo caso de estudo é um dia usual com descargas, sem a informação de falha ou não no sistema.

5.1 CASO DE ESTUDO 1:

Foi selecionado o dia 29 de Outubro de 2008, em torno do horário 13:03, em que foi registrado um desligamento de energia da LT 765 kV, no trecho entre Foz do Iguaçu

e Ivaiporã, com provável causa previamente definida como sendo descarga atmosférica. Nas Figuras 14 à 19, o período de 09:33 à 12:33 foi monitorado e a previsão para a hora seguinte foi realizada (isto é, previsão para às 13:33, sendo que o momento da falha (13:03) fica exatamente na metade do período de previsão). A Figura 14 mostra todas as tempestades elétricas que foram identificadas e acompanhadas no período mencionado pelo sistema proposto. O trajeto na cor rosa representa o caminho percorrido por cada tempestade; a seta azul é o resultado da previsão da localização da tempestade uma hora adiante e a elipse azul representa onde 50% das descargas de uma hora à frente foram previstas a incidir. Às 12:33 estavam ativas 7.037 descargas na região (isto é, incidiram 7.037 descargas na janela de tempo de 11:33 à 12:33) e 22 tempestades elétricas foram identificadas e estavam ativas no momento. Algumas tempestades apresentam um longo passado e outras encontram-se no estágio inicial de vida (círculo azul com um ponto no centro). Nota-se que uma tempestade (assinalada com um asterisco verde) se aproxima do trecho mencionado onde ocorreu a falha, por isso será analisada mais detalhadamente.

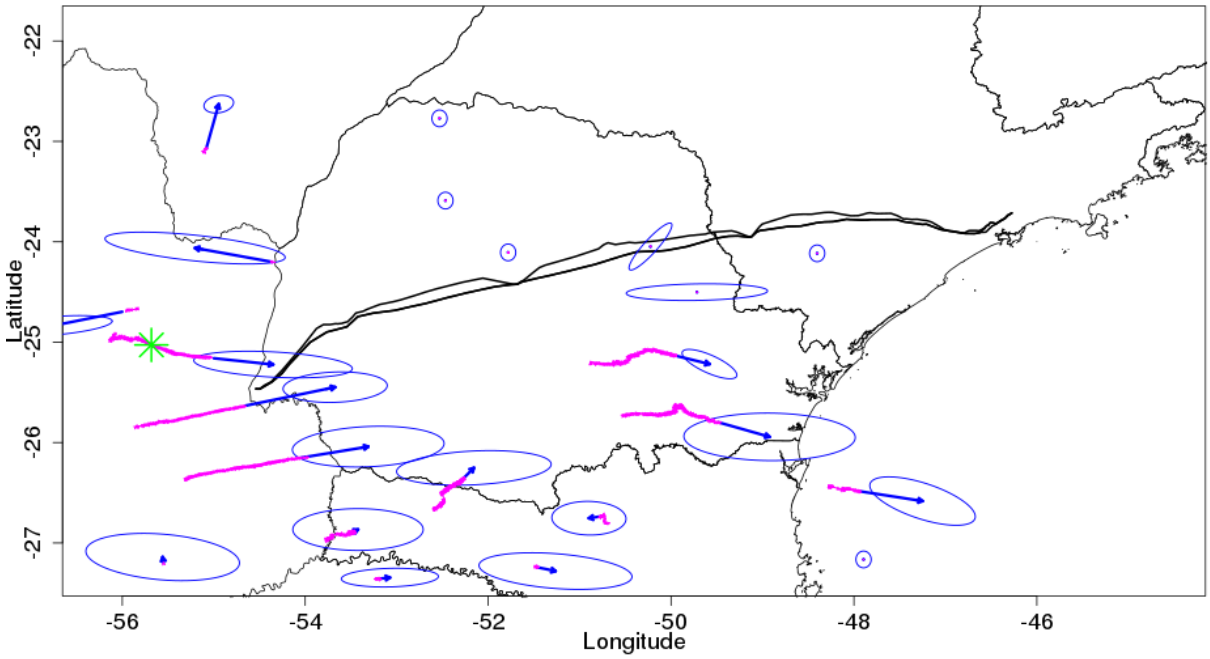


FIGURA 14: Tempestades elétricas identificadas no dia 29/10/2008 das 09:33 às 12:33 (trajetos rosas) e suas respectivas previsões uma hora à frente (setas e elipse azuis)

FONTE: A autora (2015)

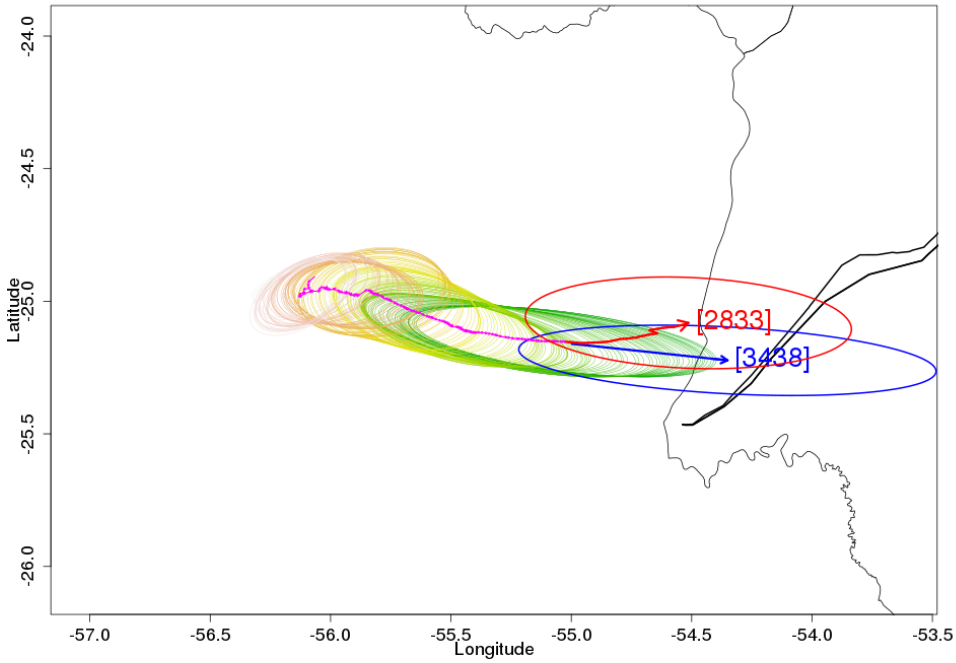


FIGURA 15: Trajetória de uma tempestade elétrica identificada na região (rosa), o trajeto real (vermelho) e sua respectiva previsão uma hora à frente (azul)

FONTE: A autora (2015)

A Figura 15 ilustra a ampliação da imagem da tempestade assinalada na figura anterior com algumas informações adicionais. As elipses em gradiente de cores (mais claras são mais antigas) mostram com 50% de certeza onde as descargas que integram a tempestade incidiram minuto a minuto. É possível notar que esta tempestade em sua fase inicial de monitoramento apresentou trajeto um pouco instável, porém em um certo momento começou a apresentar regularidade, com trajetória do oeste para o leste. Ocorreram em média 2.833 descargas/hora às 13:33 e a elipse vermelha indica a distribuição espacial das descargas futuras com 50% de confiança. Foram previstas 3.438 descargas/hora para o momento e a elipse azul ilustra a área onde estas descargas foram previstas incidir com 50% de confiança. Atenta-se que o percurso da tempestade é no sentido da linha de energia e que as descargas (elipses azul e vermelha) se distribuem nas mediações do trecho da linha em que foi registrada a falha, mostrado tanto na previsão como no real ocorrido.

A Figura 16 exibe o comportamento do número médio de descargas por hora

desta tempestade onde é possível notar o crescente aumento desta variável, chegando a mais de 2.000 descargas/hora às 12:33 (último minuto monitorado e indicado no tempo 0). Para as 13:33 foram previstas 3.438 descargas/hora (curva azul), porém o valor observado foi menor, 2833 descargas por hora (curva vermelha), conforme já relatado anteriormente. Nota-se que a curva rosa inicia aproximadamente no tempo -150, significando que esta tempestade não tem um passado de 180 minutos, e sim 150 minutos, ou seja, ela iniciou por volta de 10:00.

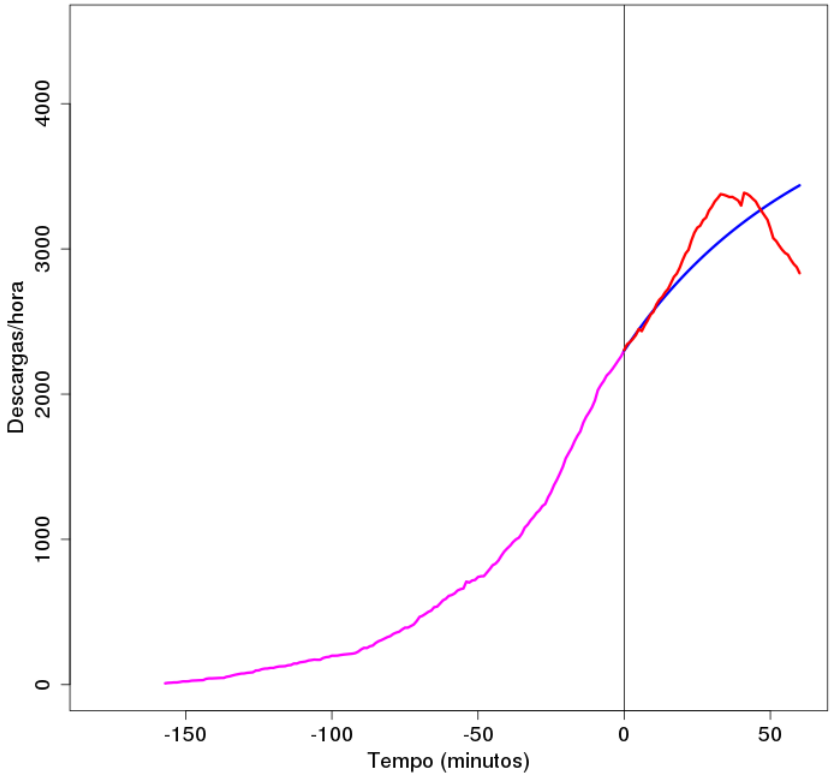


FIGURA 16: Comportamento do número médio de descargas por hora de uma tempestade elétrica (curva rosa), sua previsão (curva azul) e observação (curva vermelha) uma hora à frente

FONTE: A autora (2015)

A Figura 17 apresenta o comportamento da área da elipse de 95% de confiança da tempestade analisada, representando a abrangência espacial das descargas. Esta variável, assim como o número de descargas, apresenta uma evolução (curva rosa), atingindo quase 1 grau^2 (ou equivalentemente 100×10^3 km^2) às 12:33. Já para às 13:33, tanto a previsão (curva azul) quanto a observação (curva vermelha) foram de

aproximadamente 150×10^3 km^2, significando que as descargas futuras são mais dispersas sobre a região.

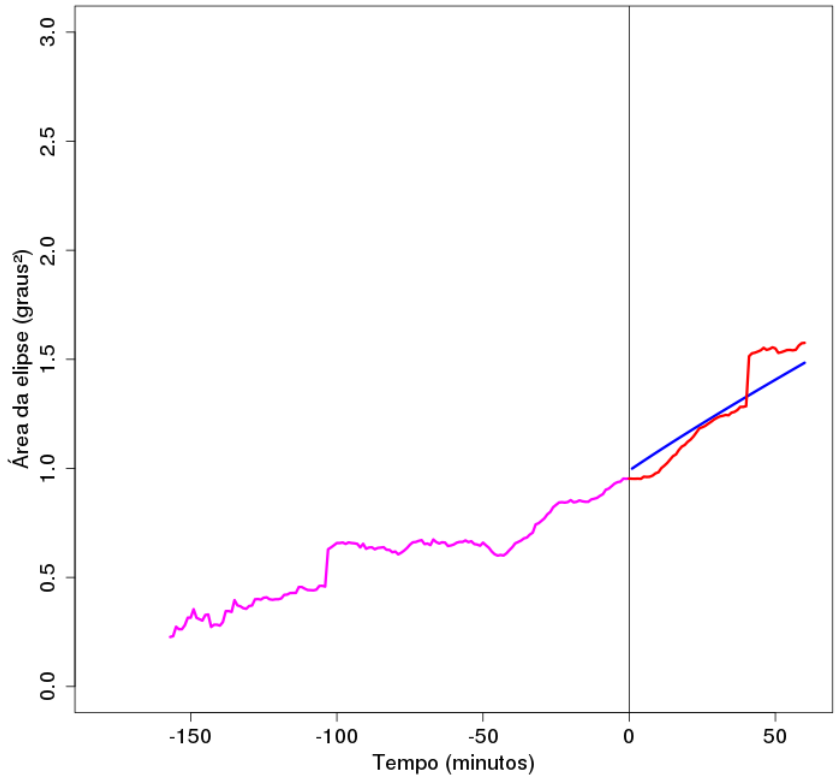


FIGURA 17: Comportamento da área da elipse de incerteza de 95% de uma tempestade elétrica (curva rosa), sua previsão (curva azul) e observação (curva vermelha) uma hora à frente

FONTE: A autora (2015)

A Figura 18 exibe o comportamento da média do valor absoluto do pico de corrente das descargas que integraram a tempestade ao longo do período analisado e sua respectiva extrapolação. Durante o período monitorado da vida desta tempestade (curva rosa), houve um máximo de pouco menos de 35 kA logo no início do rastreio, decrescendo após esse momento, até atingir valor próximo a 23 kA no último minuto de monitoramento. A previsão (curva azul) e a observação (curva vermelha) uma hora à frente resultaram em valores próximos a 20 kA. Estes resultados mostram que o valor médio do pico de corrente das descargas diminuiu mais de 10 kA com a evolução da tempestade.

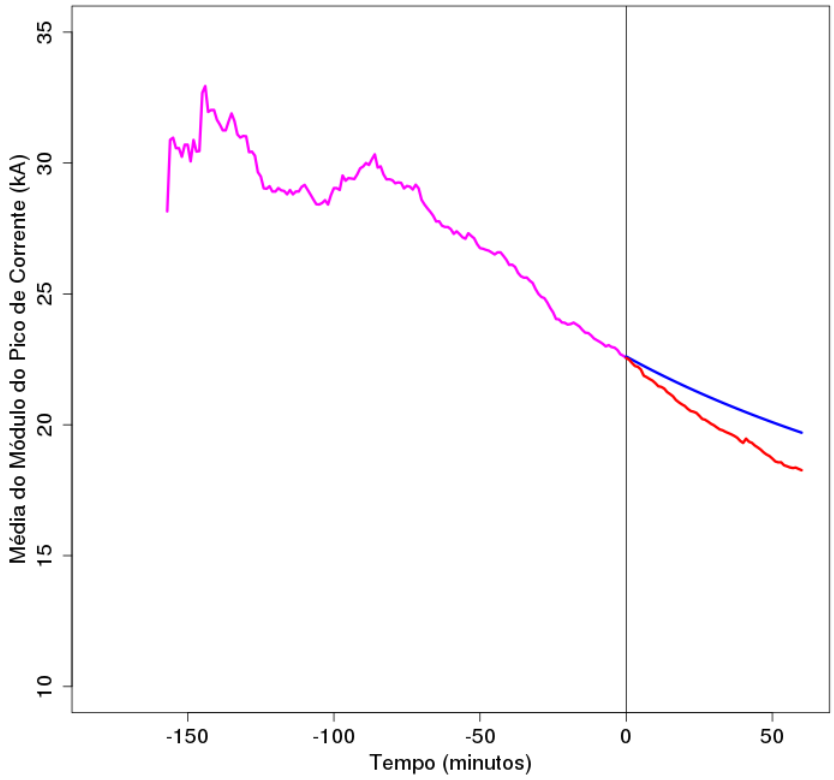


FIGURA 18: Comportamento da média do valor absoluto do pico de corrente das descargas que compõem a tempestade elétrica (curva rosa), sua previsão (curva azul) e observação (curva vermelha) uma hora à frente

FONTE: A autora (2015)

Finalmente, na Figura 19 são exibidas as mesmas informações apresentadas na Figura 14, isto é, o trajeto das 22 tempestades identificadas na região em um período de 3 horas e a previsão da localização uma hora à frente, porém com a informação do número esperado de descargas para às 13:33 em uma grade regular de 10x10 km. Analisando a mesma tempestade investigada anteriormente, nota-se que ela provoca um número bastante elevado de descargas (20 descargas para uma quadrícula de 100 km^2) para uma área contígua ao seu desenvolvimento na próxima hora. Outras duas tempestades apresentam condições similares à analisada em relação ao número esperado de descargas por quadrícula. As demais tempestades apresentaram previsões de 1 a 15 descargas a cada 100 km^2.

Por meio das Figuras 14 a 19, pode-se concluir que havia uma tempestade eletri-

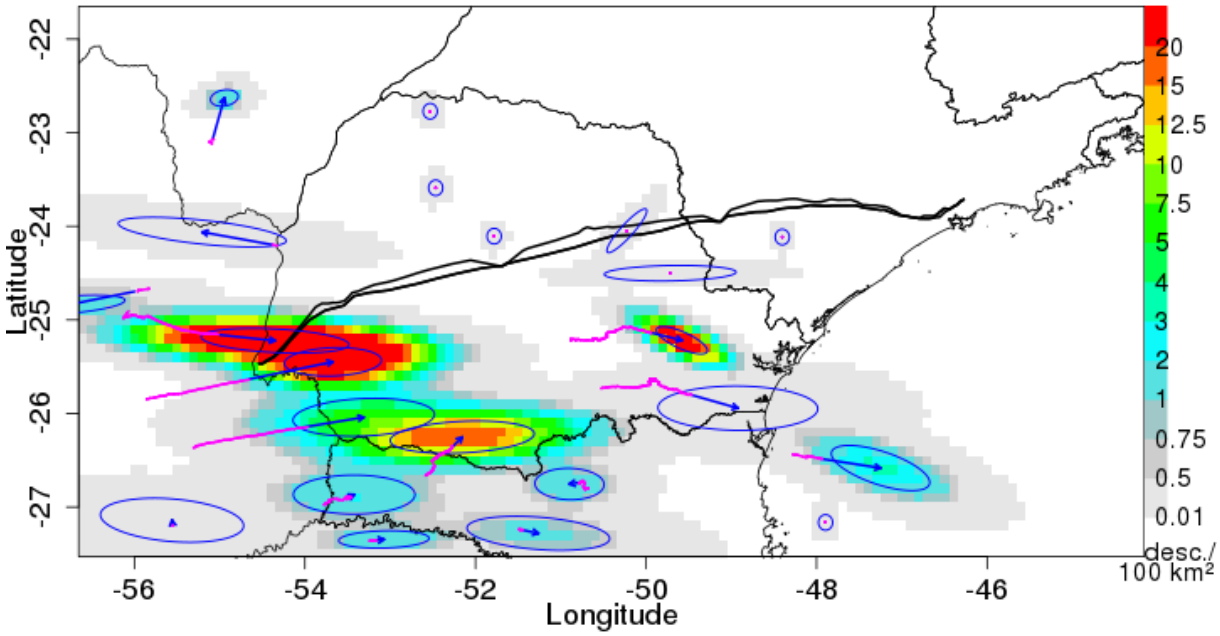


FIGURA 19: Número esperado de descargas do dia 29/10/2008 para às 13:33 em 10x10 km na região piloto

FONTE: A autora (2015)

camente ativa se desenvolvendo nas proximidades do trecho Foz do Iguaçu - Ivaiporã da LT 765 kV, e que às 13:03 (momento exato do desligamento) era esperada uma forte atividade elétrica sobre a linha. Se operado em tempo real, este sistema de identificação, monitoramento e previsão de tempestades elétricas pode representar uma poderosa ferramenta de apoio em tomadas de decisão para que falhas como esta possam ser alertadas e possíveis estratégias de operação do sistema possam ser aplicadas.

5.2 CASO DE ESTUDO 2:

O segundo caso estudado é referente ao dia 10 de Julho de 2015, onde tempestades elétricas foram identificadas e rastreadas das 11:00 até 14:00 horas e projetadas para às 15:00. Não há informação se houve ou não desligamento de energia da linha neste período. A Figura 20 ilustra 11.484 descargas organizadas em 29 tempestades elétricas ativas às 14:00, com suas previsões para às 15:00.

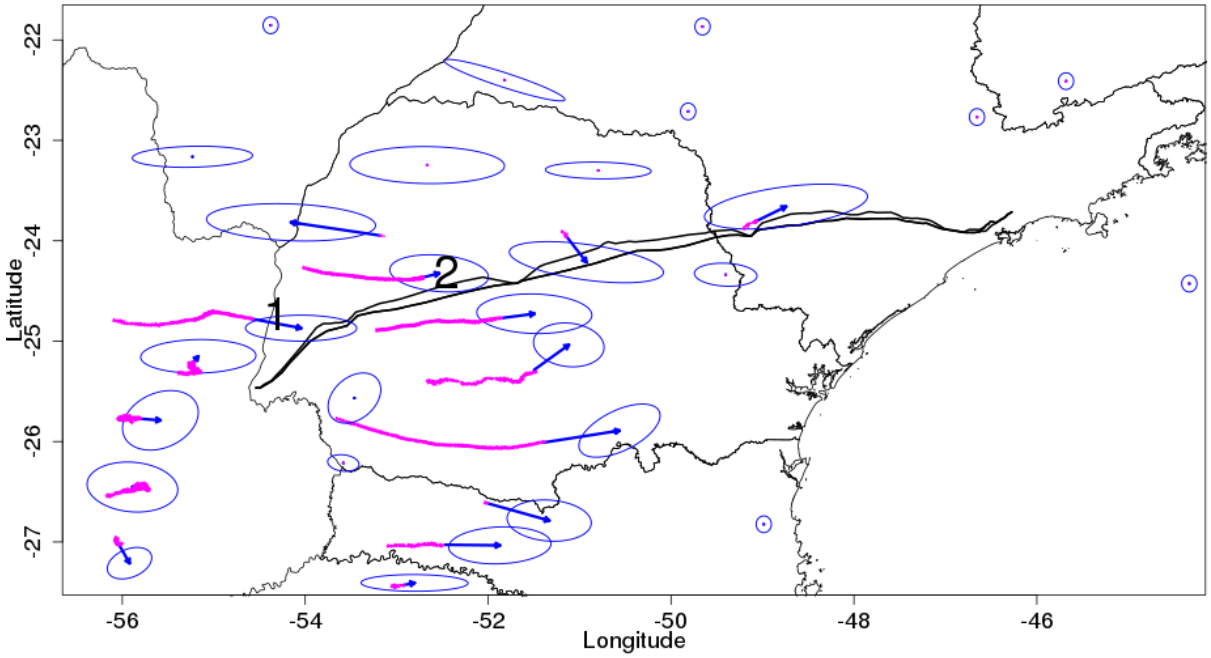
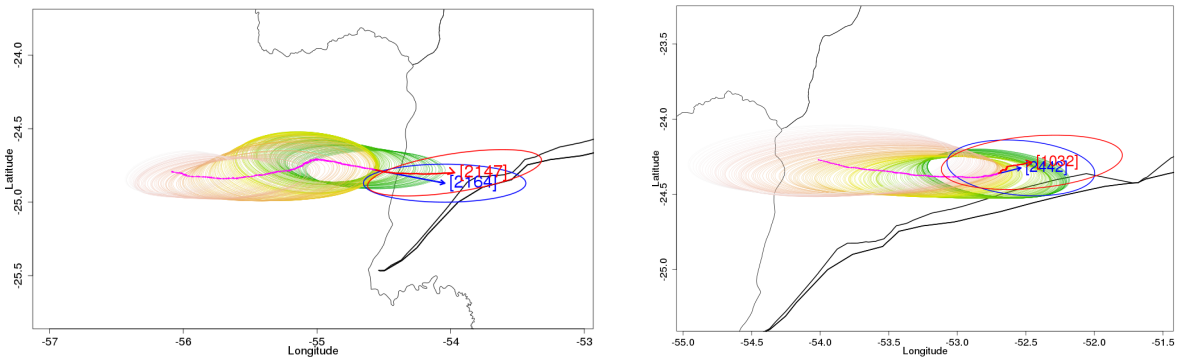


FIGURA 20: Tempestades elétricas identificadas no dia 10/07/2015 das 11:00 às 14:00 (trajetos rosas) e suas respectivas previsões uma hora à frente (setas e elipses azuis)

FONTE: A autora (2015)

Duas tempestades elétricas (n° 1 e n° 2 na Figura 20) que se desenvolveram nos arredores da LT 765 kV, mais especificamente no trecho entre Foz do Iguaçu e Ivaiporã, serão investigadas mais detalhadamente a seguir. A Figura 21 ilustra a ampliação do trajeto da tempestade n° 1 (Figura 21a) e da tempestade n° 2 (Figura 21b). As trajetórias das duas tempestades são bastante regulares (rosa) e os trajetos previstos (azuis) e observados (vermelhos) são próximos nos dois casos, assim como a representação espacial das descargas previstas e observadas (elipses azuis e vermelhas respectivamente). Na Figura 21a, as elipses em cores degradê revelam que o espalhamento das descargas foi menor na fase inicial do que na fase de maturação da tempestade n° 1 no período monitorado (as elipses em rosa claro são menores do que as elipses em amarelo), porém adiante voltam a se tornar mais agrupadas (elipses verdes). Já a Figura 21b indica um espalhamento maior das descargas no início do monitoramento da tempestade n° 2. A previsão do número médio de descargas por hora (2.164 descargas/hora) da tempestade n° 1 foi bastante próximo ao real ocorrido (2.147 descargas/hora), porém para a tempestade n° 2 houve uma superestimação

deste valor (ocorreram 1.032 e o sistema previu 2.442 descargas/hora).



(A) Tempestade n° 1

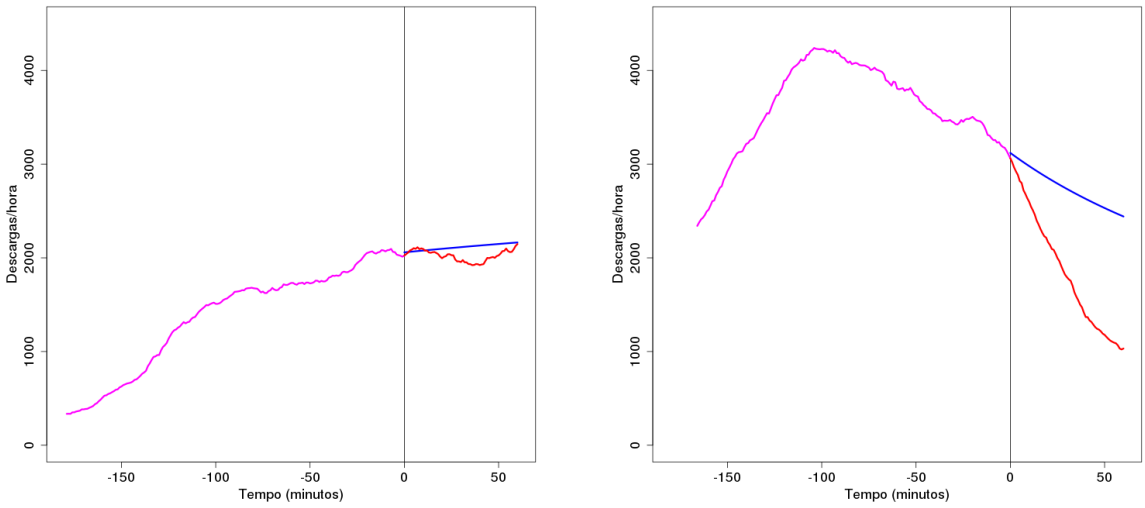
(B) Tempestade n° 2

FIGURA 21: Trajetórias de duas tempestades elétricas identificadas na região (rosas), trajetos reais (vermelhos) e suas respectivas previsões uma hora à frente (azuis)

FONTE: A autora (2015)

A Figura 22 ilustra os resultados do monitoramento do número de descargas das 11:00 às 14:00 (curvas rosas), previsões (curvas azuis) e observações (curvas vermelhas) das duas tempestades analisadas. A Figura 22a exibe as curvas monitorada, prevista e observada da tempestade n° 1, onde nota-se que durante o monitoramento o número médio de descargas por hora ascendeu até aproximadamente 2000, e a previsão foi próxima do real valor observado uma hora adiante. A Figura 22b, referente a tempestade n° 2, realça um pico de mais de 4000 descargas/hora no tempo -100 aproximadamente (80 minutos após o início do monitoramento) e após esse pico, esta variável começa a cair, e a previsão estimada foi bastante superior ao real valor observado, conforme já mencionado e indicado na Figura 21b. Comparando as Figuras 22a e 22b, nota-se a maior severidade da tempestade n° 2 do que da tempestade n° 1 em relação ao número médio de descargas durante o tempo monitorado.

A Figura 23 exibe o comportamento a área da elipse de incerteza das duas tempestades analisadas. A Figura 23a aponta que a tempestade n° 1, no início do tempo analisado, se iniciou pouco dispersa em relação as descargas que a estruturam, tendo um espalhamento maior na metade do período monitorado e concentrando-se mais ao final das 3 horas acompanhadas. A Figura 23b indica que as descargas são mais dis-



(A) Tempestade n° 1

(B) Tempestade n° 2

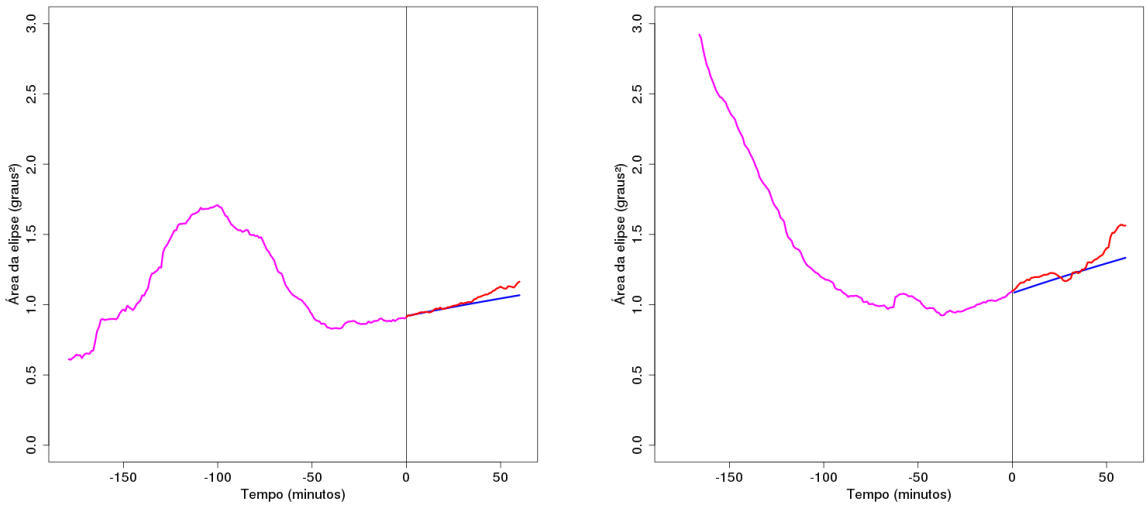
FIGURA 22: Comportamento do número médio de descargas por hora de duas tempestades elétricas (curvas rosas), suas previsões (curvas azuis) e observações (curvas vermelhas) uma hora à frente

FONTE: A autora (2015)

persas na tempestade n° 2 no início e tendem a se concentrar quando a tempestade é bem comportada e está no seu estágio maduro. A tempestade n° 2 apresenta, no geral, área de elipse maior do que a tempestade n° 1. Ambas as tempestades tem previsão próxima à observação uma hora adiante.

A Figura 24 apresenta o comportamento da média do valor absoluto do pico de corrente das duas tempestades estudadas neste caso de estudo. O comportamento desta variável é similar nas tempestades n° 1 (Figura 24a) e n° 2 (Figura 24): no início do monitoramento é maior e decresce ao passar do tempo, porém a tempestade n° 1 tem média um pouco superior que a tempestade n° 2. Ambas as previsões foram muito boas em comparação com os valores observados.

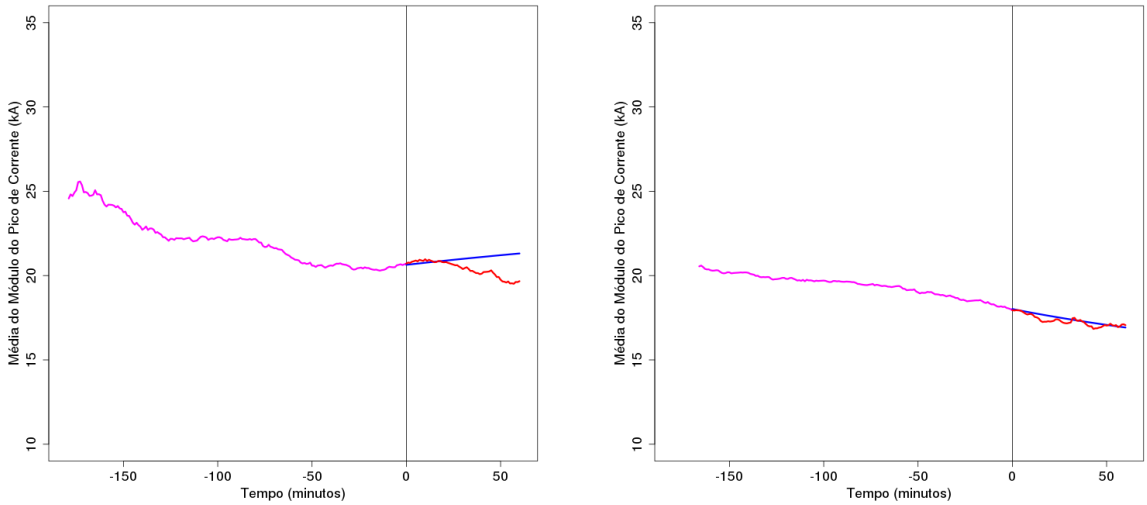
O número esperado de descargas por quadrícula 10x10 km uma hora adiante para as 29 tempestades identificadas é apresentado na Figura 25. Percebe-se que há tempestades que cursam organizadamente no sentido sudoeste para nordeste na região piloto, algumas cruzando a LT 765 kV, originando previsões de mais de 20 descargas por quadrícula para áreas adjacentes às tempestades. Não se sabe se



(A) Tempestade n° 1 (B) Tempestade n° 2

FIGURA 23: Comportamento da área da elipse de incerteza de 95% de duas tempestades elétricas (curvas rosas), sua previsões (curvas azuis) e observações (curvas vermelhas) uma hora à frente

FONTE: A autora (2015)



(A) Tempestade n° 1 (B) Tempestade n° 2

FIGURA 24: Comportamento da média do valor absoluto do pico de corrente das descargas que compõem duas tempestades elétricas (curvas rosas), sua previsões (curvas azuis) e observações (curvas vermelhas) uma hora à frente

FONTE: A autora (2015)

houve desligamento na linha neste período analisado.

Apenas a fim de verificação, a Figura 26 apresenta dados de refletividade medidos

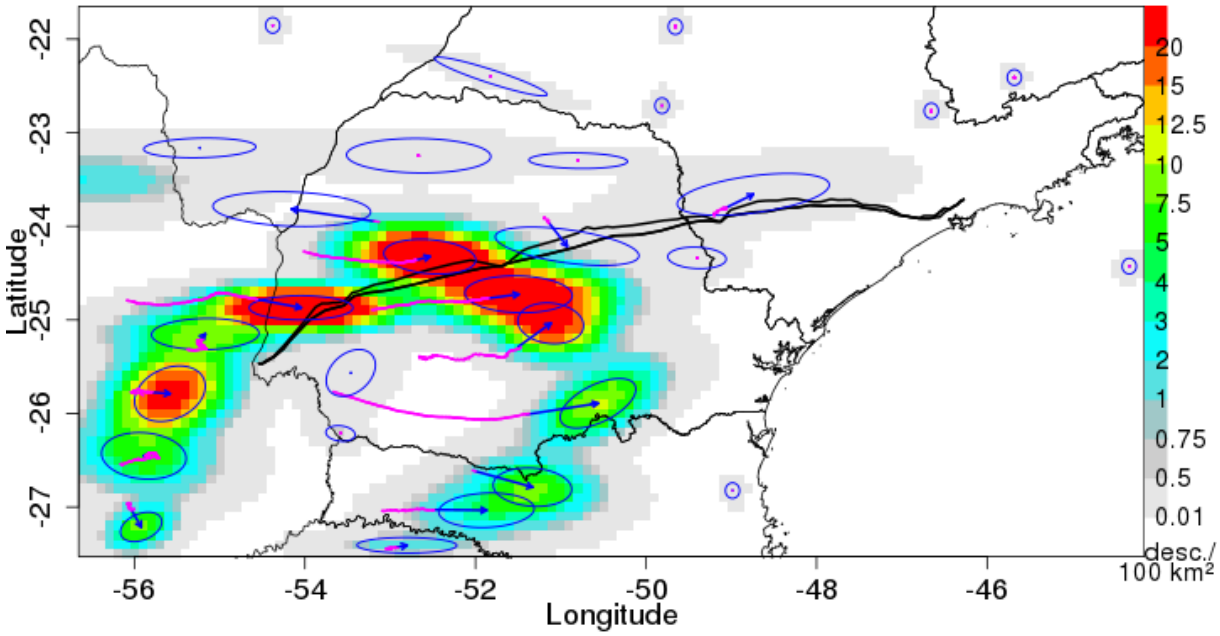


FIGURA 25: Número esperado de descargas do dia 10/07/2015 para às 14:00 em 10x10 km na região piloto

FONTE: A autora (2015)

pelo radar meteorológico operado pelo SIMEPAR instalado na cidade de Teixeira Soares/PR, com alcance de 480 km. As imagens são dia analisado, isto é, 10/07/2015, de hora em hora para o período de monitoramento (11:00, 12:00, 13:00 e 14:00) e previsão (15:00), respectivamente. Nota-se, pelas sucessivas imagens de radar, uma grande célula de tempestade deslocando-se do oeste para o leste da região.

Analisando a imagem do radar (Figura 26) com a informação adicional das descargas dos últimos 5 minutos no momento da previsão (15:00), observa-se que ocorre uma grande quantidade de descargas na parte frontal da linha de instabilidade. A previsão realizada pelo sistema proposto (Figuras 20 e 25) é coerente com a imagem do radar e com as informações das descargas incidentes.

Mais ilustrações do sistema de identificação, monitoramento e previsão de tempestades elétricas para outros períodos podem ser encontrados no Apêndice A.

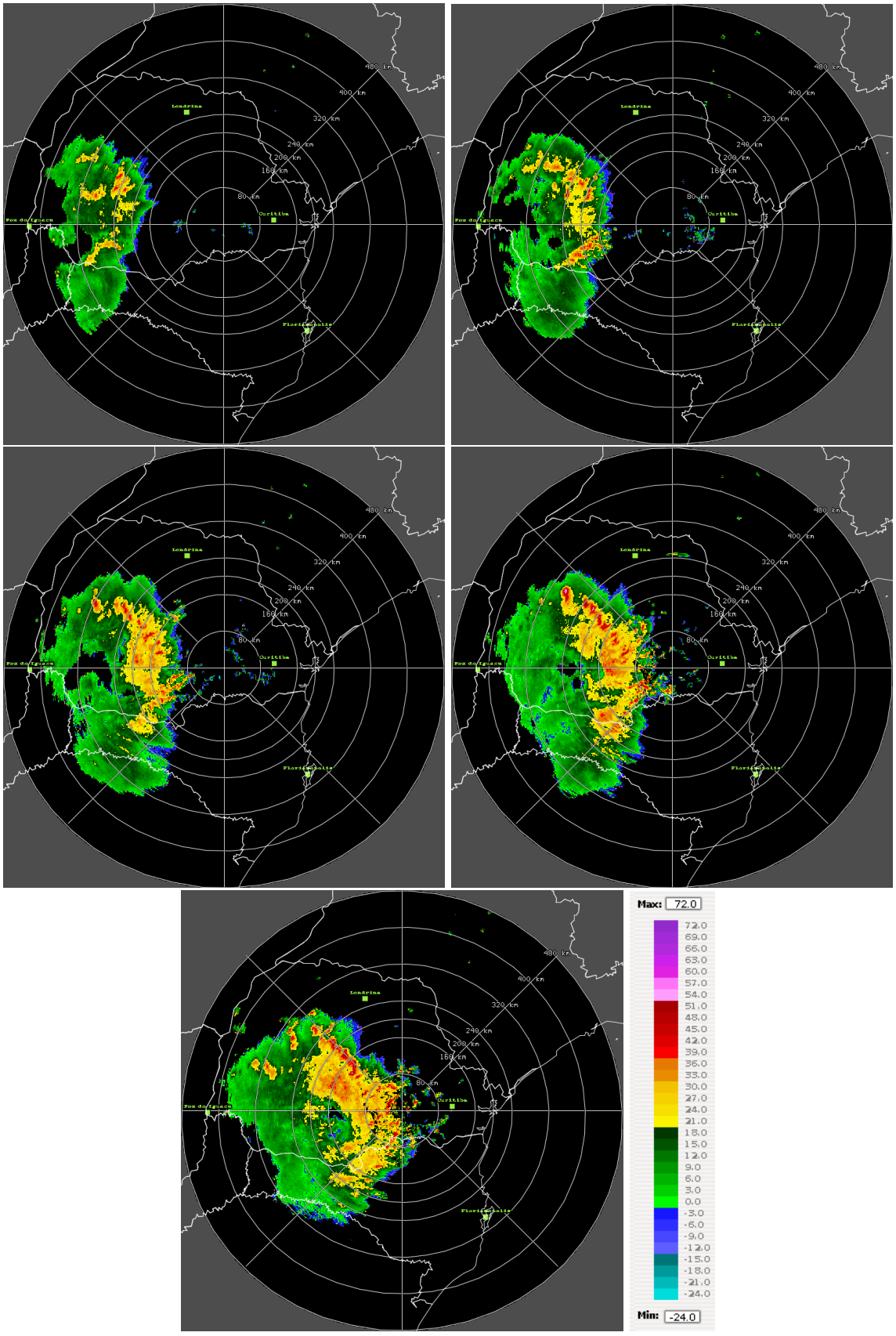


FIGURA 26: Imagens de radar do dia 10/07/2015 às 11:00, 12:00, 13:00, 14:00 e 15:00, respectivamente

FONTE: A autora (2015)

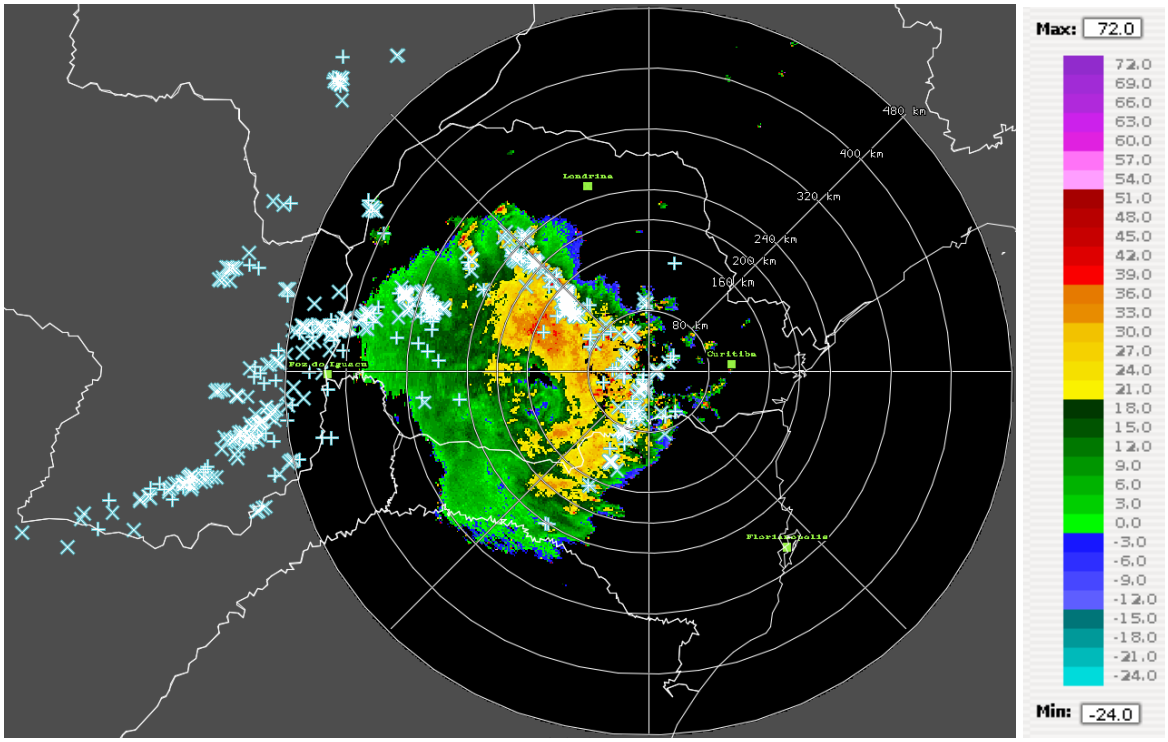


FIGURA 27: Imagem de radar com a informação das descargas do dia 10/07/2015 às 15:00
FONTE: A autora (2015)

5.3 CASO DE ESTUDO 1 $\times$ CASO DE ESTUDO 2

Apesar de não ser possível generalizar para todas as situações e que os exemplos apresentados foram apenas para ilustração dos resultados do sistema proposto, dá para se ter uma ideia de magnitude de uma tempestade apontada como possível causadora de falha no sistema elétrico e de uma tempestade sem informação de falha por meio da comparação entre o primeiro e segundo caso de estudo apresentado nesta pesquisa.

Comparando as Figuras 14 e 20, nota-se que há mais tempestades elétricas nas proximidades da LT 765 kV no caso de não informação de falha do que na situação que ocorreu falha na linha de energia.

Ao analisar a Figura 16, atenta-se que o número médio de descargas por hora está em pleno crescimento no tempo $t = 0$, enquanto que a Figura 22a mostra que a tempestade n° 1 não tem crescimento acentuado como na situação de falha e a Figura

22b aponta que a tempestade n° 2 teve um valor extremo maior que na situação de falha, porém ao final do período de monitoramento ($t = 0$) está decrescendo.

Em relação a distribuição espacial das descargas dentro de uma tempestade elétrica, na situação de falha (Figura 17) o espalhamento está aumentando ao final do monitoramento, enquanto que as duas tempestades analisadas no segundo caso (Figura 23) apresentam decréscimo da área da elipse de incerteza, ao término do monitoramento.

A média do valor absoluto do pico de corrente das descargas, tanto no caso 1 quanto no caso 2, apresentam comportamento similar: decrescem no decorrer do rastreio das tempestades analisadas. Porém no caso de falha analisado (Figura 18), no geral esta variável apresenta valor superior ao caso de não informação de falha na linha de energia (Figura 24).

O grande diferencial do sistema proposto em relação aos demais existentes no mesmo âmbito (brevemente descritos no início do Capítulo 4) é justamente o fato deste prever o pico de corrente das descargas que compõem uma tempestade e este conhecimento pode representar uma informação diferencial no alerta de riscos à áreas afetadas por intensa atividade elétrica, já que sabe-se que valores altos de pico de corrente são mais propícios a causar falhas no sistema elétrico.

5.4 OBSERVAÇÕES FINAIS SOBRE O SISTEMA PROPOSTO

Um requisito muito importante e fundamental do sistema é que ele seja capaz de capturar o ciclo de vida das tempestades elétricas. Interrupções inadequadas e não realistas das tempestades podem acarretar em perturbações nas variáveis monitoradas. Para inspecionar essa particularidade dos sistema proposto, um bom indicador para tal é analisar o número de descargas das tempestades, pois geralmente no início esse número é pequeno, cresce na fase de maturação e cai ao final da tempestade. A Figura 28 ilustra o número médio de descargas das 29 tempestades identificadas no caso de estudo 2 (Seção 5.2) no período de 180 minutos de monitoramento. É

possível notar que algumas tempestades estão se desenvolvendo (variável começa baixa e vai crescendo ao longo do tempo) e outras tempestades já estão findando (no tempo $t = 0$ a variável está diminuindo). Esse ciclo (crescimento e decaimento da variável) é um bom indicativo de que o sistema tem a capacidade de representação das tempestades elétricas ativas ao longo do tempo.

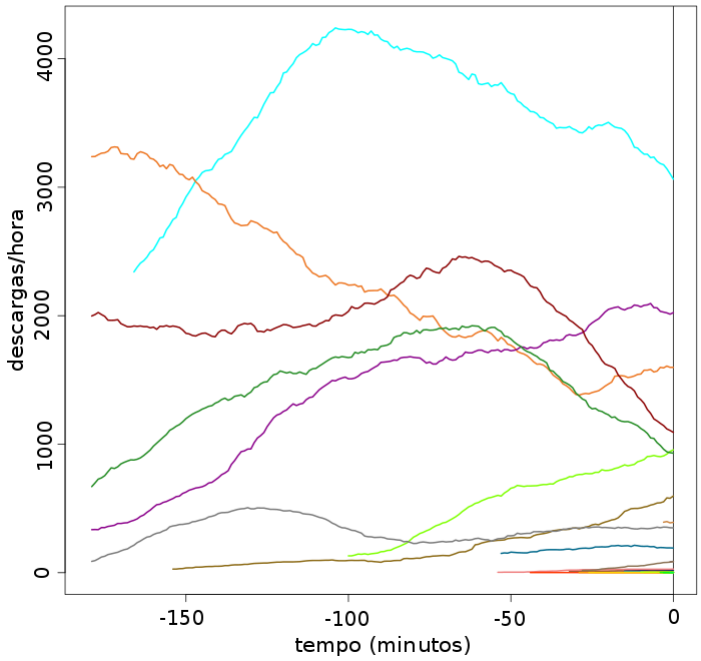
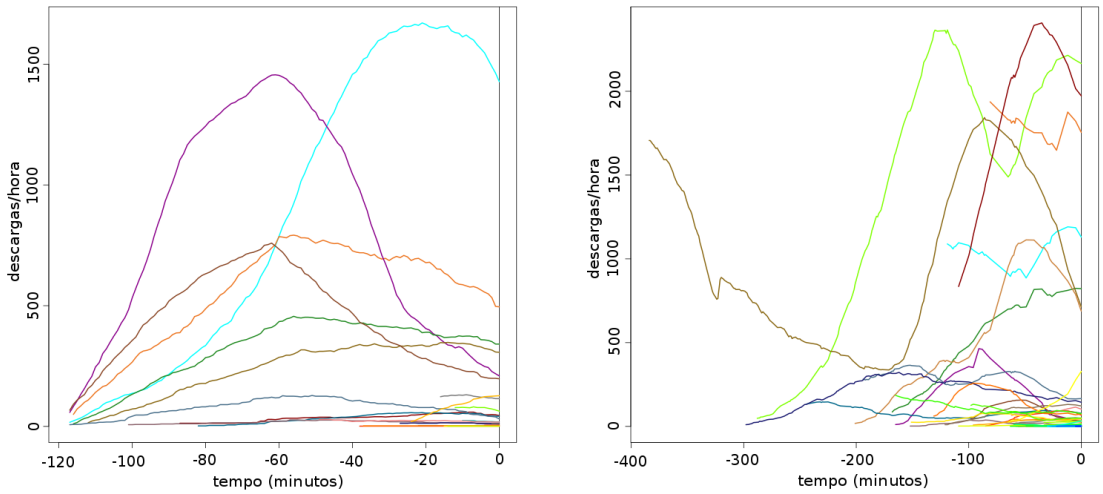


FIGURA 28: Número médio de descargas por hora das 29 tempestades elétricas identificadas no segundo caso estudado, isto é, em 10/07/2015 das 11:00 às 14:00

FONTE: A autora (2015)

Outros exemplos mais compreensíveis para visualizar os ciclos de vidas das tempestades elétricas monitoradas pelo sistema proposto são apresentados na Figura 29.



(A) 08:22 às 11:22 de 30/09/2015 (B) 06:17 às 09:17 de 08/10/15

FIGURA 29: Comportamento do número médio de descargas por hora de tempestades elétricas de outros dois períodos monitorados com intensa atividade elétrica

FONTE: A autora (2015)

6 CONCLUSÕES

Nesta pesquisa foi apresentado um novo sistema de detecção, monitoramento e previsão de tempestades elétricas cuja base de dados são apenas informações de descargas atmosféricas de um período de poucas horas. Os métodos matemáticos que idealizam o sistema proposto não necessitam de um grande conjunto de dados para realizar os cálculos, como é o caso de uma rede neural que demanda uma grande quantidade de dados em seu treinamento. Este fato favoreceu a calibração do sistema, com baixo custo computacional.

Na etapa de identificação e monitoramento, fez-se uso de uma técnica de clusterização chamada *Convergent Data Sharpening* que apontou centros de atividades elétricas e, juntamente com conexões espaciais/temporais, caracterizou as chamadas tempestades elétricas. A utilização do método *Convergent Data Sharpening* foi fundamental para o êxito da aplicação, pois por meio de sua característica de reduzir o conjunto de dados para modos locais, permitiu um melhor acompanhamento do trajeto das tempestades elétricas, seguindo o comportamento prevalecente dos eventos meteorológicos da região.

Na etapa de previsão, fez-se uso da técnica de extrapolação de dados para prever, uma hora à frente, atributos das tempestades elétricas identificadas na etapa anterior.

Para validar e calibrar o sistema proposto, um problema de otimização foi desenvolvido a fim de encontrar um valor para o parâmetro do método de clusterização que aliasse bom monitoramento e previsibilidade, representando as tempestades elétricas em uma grade regular e comparando valores previstos e observados de atributos por quadrícula, tanto na etapa de monitoramento quanto na previsão. Após ajustado, o novo sistema pôde ser empregado na região piloto, visando o acompanhamento e vigilância de eventos com intensa atividade elétrica, particularmente nas proximidades

da LT 765 kV, a linha de transmissão de tensão mais elevada do Brasil.

A aplicação do sistema proposto em situações reais permitiu a visualização dos resultados de períodos com intensa atividade elétrica nas proximidades da linha de transmissão presente na região piloto. Foi possível acompanhar visualmente a trajetória de todas as tempestades elétricas ativas por um período de 3 horas (tempo máximo de memória do sistema apresentado), e selecionada uma tempestade, observou-se alguns de seus atributos, tais como o número médio de descargas e média do valor absoluto do pico de corrente de descargas por tempestade. Além do monitoramento das características do passado das tempestades elétricas, também foi possível acompanhar a previsão destas mesmas características uma hora à frente, possibilitando uma percepção da severidade e estágio de vida de uma tempestade.

Duas situações distintas foram discutidas na pesquisa: a primeira foi a análise de uma tempestade apontada como a possível causadora de uma falha em um trecho da linha de energia LT 765 kV e a segunda foi o estudo de duas tempestades, também nas proximidades da mesma linha, porém sem a informação de falha da linha. O resultado do monitoramento e previsão mostrou que, para os casos específicos analisados, a variável relacionada ao pico de corrente das descargas pode ter sido o principal diferencial entre as tempestades nas situações de falha e não informação de falha (casos de estudo 1 e 2 respectivamente). O sistema apresentado, sendo capaz de realizar previsões do pico de corrente das descargas dentro de tempestades elétricas, pode representar uma ferramenta importante para o alerta de riscos em áreas que englobam linhas de transmissão de energia.

De um modo geral, a técnica de clusterização se mostrou uma ferramenta com alto potencial para identificar tempestades elétricas, bem como a técnica de extrapolação de dados para prever atributos das mesmas. Em relação ao sistema proposto, os resultados tanto no monitoramento quanto na previsão se mostraram adequados e coerentes com os eventos meteorológicos atuantes na região de estudo.

A utilização de informações de descargas atmosféricas apenas (não utiliza dados

de radares e satélites, por exemplo), sem a necessidade de um grande volume de dados, além da previsão da característica elétrica da tempestade (pico de corrente), diferencia o sistema proposto dos demais existentes.

6.1 TRABALHOS FUTUROS

Apesar da satisfatoriedade com o sistema proposto, muitos aspectos podem ser aprimorados, tais como:

- Ampliação da região de aplicação, abrangendo mais áreas de interesse onde descargas atmosféricas causam prejuízos;
- Dividir a região de estudo em subregiões de acordo com a localidade ou de acordo com os próprios dados de descargas e utilizar um parâmetro h do método de clusterização para cada uma destas subregiões a fim de se obter um agrupamento mais coerente, podendo-se levar em conta características meteorológicas locais;
- Utilizar modelos numéricos mais efetivos para previsão do número de descargas por tempestades;
- Incorporar outros tipos de descargas atmosféricas, tais com descargas intranuvens, que em conjunto com as descargas nuvem-solo, favoreçam a identificação mais realista e exata de tempestades elétricas;
- Elaborar uma interface gráfica para visualização na aplicação em tempo real (modo operacional) do sistema e criar um índice que envolva diversos atributos de uma tempestade elétrica indicando se é ou não uma ameaça em certas áreas de riscos;
- Agregar dados de outras fontes, tais como radar, satélite e vento telemedido.

REFERÊNCIAS

ALBUQUERQUE, M. A. de. **Análise de Agrupamento Hierárquica e Incremental - Estudo de Caso em Ciências Florestais**. Dissertação (Mestrado) — Universidade Federal Rural de Pernambuco, 2013.

ANKERST, M. *et al.* OPTICS: Ordering points to identify the clustering structure. In: **Proceedings of the ACMSIGMOD Conference on Management of Data**. Philadelphia: ACM, 1999. p. 49–60.

BABUSKA, R. **Fuzzy Modeling for Control**. New York: Springer, 1998.

BENETI, C. A. A. **Caracterização Hidrodinâmica e Elétrica de Sistemas Convectivos de Mesoescala**. Tese (Doutorado) — Instituto de Astronomia, Geofísica e Ciências Atmosféricas, Universidade de São Paulo, 2012.

BERGER, K.; ANDERSON, R. B.; KRÖNINGER, H. Parameters of lightning flashes. **Electra**, v. 41, p. 23–37, 1975.

BETZ, H. D. *et al.* Cell-tracking with lightning data from LINET. **Advances in Geosciences**, v. 17, p. 55–61, 2008.

BIONDI NETO, L. *et al.* Minicurso de sistema especialista nebuloso. In: **XXXVII Simpósio Brasileiro de de Pesquisa Operacional**. Goiânia: [s.n.], 2006.

BONATO, J. V. R. **Clusterização de Dados Meteorológicos para Comparação de Técnicas de Nowcasting**. Dissertação (Mestrado) — Universidade Federal do Paraná, 2014.

BONELLI, P.; MARCACCI, P. Thunderstorm nowcasting by means of lightning and radar data: algorithms and applications in northern Italy. **Natural Hazards and Earth System Sciences**, v. 8, p. 1187–1198, 2008.

BORA, D. J.; GUPTA, A. K. A comparative study between fuzzy clustering algorithm and hard clustering algorithm. **International Journal of Computer Trends and Technology**, v. 10, n. 2, p. 108–113, 2014.

BOSCARIOLI, C. **Análise de Agrupamentos baseada na Topologia dos Dados e em Mapas Auto-organizáveis**. Tese (Doutorado) — Escola Politécnica da Universidade de São Paulo, 2008.

BRAGA, J. L. P. V. **Desempenho de Linhas de Distribuição Frente a Descargas Atmosféricas: Estudo e Implementaçã do Guia IEEE Std 1410**. Dissertação (Mestrado) — Universidade Federal de Minas Gerais, 2009.

CARLANTONIO, L. M. di. **Novas Metodologias para Clusterização de Dados**. Tese (Doutorado) — Universidade Federal do Rio de Janeiro, 2001.

CAVALCANTI JÚNIOR, N. L. **Clusterização baseada em algoritmos fuzzy**. Dissertação (Mestrado) — Centro de Informática, Universidade Federal de Pernambuco, 2006.

CECIL, D. J.; BUECHLER, D. E.; BLAKESLEE, R. J. Gridded lightning climatology from TRMM-LIS and OTD: Dataset description. **Atmospheric Research**, v. 135–136, p. 404–414, 2014.

CHAPRA, S. C.; CANALE, R. P. **Métodos Numéricos para Engenharia**. 5. ed. São Paulo: McGraw-Hill, 2008.

CHOI, E.; HALL, P. Data sharpening as a prelude to density estimation. **Biometrika**, v. 86, p. 941–947, 1999.

CHOWDHURI, P. Estimation of flashover rates of overhead power distribution lines by lightning strokes to nearby ground. **IEEE Transactions on Power Delivery**, v. 4, n. 3, p. 1982–1989, 1989.

COPEL. **Descargas Atmosféricas**. 2014. Disponível em: http://www.copel.com, acesso em: 28 Julho 2015.

COUTO, E. C.; DUARTE, J. V.; SOARES, M. R. Análise da taxa de falha de transformadores aéreos de distribuição. **Revista Eletricidade Moderna**, p. 54–74, 1995.

CUMMINS, K. L.; KRIDER, E. P.; MALONE, M. D. The U.S. national lightning detection network and applications of cloud-to-ground lightning data by electric power utilities. **IEEE Transactions on Electromagnetic Compatibility**, v. 40, n. 4, p. 465–480, 1998.

DIENDORFER, G.; PISTAUER, A. Lightning performance of high voltage power lines. In: **International Lightning Detection Conference**. Arizona: [s.n.], 2010.

DIENDORFER, G.; SCHULZ, W. Ground flash density and lightning exposure of power transmission lines. In: **Power Tech Conference Proceedings**. Bologna: IEEE, 2003. v. 3.

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

DÖRING, C.; LESOT, M. J.; KRUSE, R. Data analysis with fuzzy clustering methods. **Computational Statistics and Data Analysis**, v. 51, p. 192–214, 2006.

ESTER, M. *et al.* A density-based algorithm for discovering clusters in large spatial databases with noise. In: **International Conference on Knowledge Discovery and Data Mining (KDD-96)**. [S.l.: s.n.], 1996.

FRALEY, C.; RAFTERY, A. E. Model-based methods of classification: Using the mclust software in chemometrics. **Journal of Statistical Software**, v. 18, 2007.

FRALEY, C. *et al.* Mclust version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation. **Technical Report**, n. 597, 2012. Disponível em: <http://www.stat.washington.edu/research/reports/2012/tr597.pdf>.

FREITAS, J. C. *et al.* Análise de agrupamentos na identificação de regiões homogêneas de Índices climáticos no estado da paraíba, pb – brasil. **Revista Brasileira de Geografia Física**, v. 6, n. 4, p. 732–748, 2013.

GILAT, A.; SUBRAMANIAM, V. **Métodos Numéricos para Engenheiros e Cientistas: Uma Introdução com Aplicações Usando o MATLAB**. Porto Alegre: Bookman, 2008.

GOLDE, R. H. **Lightning, Physics of lightning**. New York: Academic press, 1997.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. On clustering validation techniques. **Journal of Intelligent Information Systems**, v. 17, p. 107–145, 2001.

HEIDLER, F. *et al.* Parameters of lightning current given in IEC 62305 – background, experience and outlook. In: 29^{th} **International Conference on Lightning Protection**. Uppsala: [s.n.], 2008.

HERING, A. M. *et al.* Nowcasting thunderstorms in the alpine region using a radar based adaptive thresholding scheme. **Proceedings of ERAD**, v. 10, n. 6, p. 1–6, 2004.

INOUYE, R. T. **Impacto da Assimilação de Dados Observacionais no Prognóstico de Tempo com o Modelo WRF**. Dissertação (Mestrado) — Universidade Federal do Paraná, 2014.

INPE. **Instituto Nacional de Pesquisas Espaciais**. 2015. Disponível em: http://www.inpe.br, acesso em: 15 Maio 2015.

ITAIPU. **Integração ao Sistema Brasileiro**. 2013. Disponível em: http://www.itaipu.gov.br/energia/integracao-ao-sistema-brasileiro, acesso em: 04 Nov 2013.

JOLLIFFE, I. T.; STEPHENSON, D. B. **Forecast Verification: A Practitioner's Guide in Atmospheric Science**. England: Wiley, 2003.

KING, K. **Primer on effects of lightning on electrical T&D**. 2003. Electric Light & Power.

KLEINA, M.; MATIOLI, L. C.; LEITE, E. A. Análise da intensidade do pico de corrente de descargas elétricas associadas à tempestades identificadas por técnicas de clusterização. **Proceeding Series of the Brazilian Society of Computational and Applied Mathematics**, v. 2, 2014.

KLEINA, M.; MATIOLI, L. C.; LEITE, E. A. **IDENTIFICAÇÃO, MONITORAMENTO E PREVISÃO DE TEMPESTADES ELÉTRICAS UTILIZANDO MÉTODOS NUMÉRICOS**. 2015. Artigo aceito para publicação pela revista Boletim de Ciências Geodésicas.

KLEINA, M. *et al.* **Identificação e Análise das Características de Tempestades Elétricas Utilizando Métodos Numéricos de Agrupamento**. 2015. Artigo aceito pela revista Anuário do Instituto de Geociências, com perspectiva de publicação no v. 38 - 2.

KOHONEN, T. **Self-Organizing Maps**. 3. ed. Berlin: Springer, 2001.

LEITE, E. A. *et al.* Metodologia para análise e correlação entre desligamento e incidência de descargas atmosféricas. In: **XX Seminário Nacional de Produção e Transmissão de Energia Elétrica**. Recife, Pernambuco: [s.n.], 2009.

LOPES, T. J. da S. **Extensão de um algoritmo de subspace clustering para a organização tridimensional de dados de expressão gênica**. Dissertação (Mestrado) — Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo, 2006.

MACGORMAN, D. R.; RUST, W. D. **The Electrical Nature of Storms**. New York: Oxford University Press, 1998.

MUELLER, C. *et al.* NCAR auto-nowcast system. **Weather and Forecasting**, v. 18, p. 545–561, 2003.

NUCCI, C. A. A survey on cigré and IEEE procedures for the estimation of the lightning performance of overhead transmission and distribution lines. In: **Asia-Pacific Symposium on Electromagnetic Compatibility**. Beijing: IEEE, 2010. p. 1124–1133.

OLIVEIRA, T. B. S. de. **Clusterização de dados utilizando técnicas de redes complexas e computação bioinspirada**. Dissertação (Mestrado) — Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo, 2008.

QIU, B.-Z.; ZHANG, X.-Z.; SHEN, J.-Y. Grid-based clustering algorithm for multi-density. In: **Fourth International Conference on Machine Learning and Cybernetics**. Guangzhou: IEEE, 2005. v. 3.

QUINTAL, G. M. da C. C. **Análise de clusters aplicada ao Sucesso/Insucesso em Matemática**. Dissertação (Mestrado) — Departamento de Matemática e Engenharias, Universidade da Madeira, 2006.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2012. ISBN 3-900051-07-0. Disponível em: <http://www.R-project.org/>.

RAKOV, V. A.; UMAN, M. A. **Lightning, Physics and Effects**. New York: Cambridge University Press, 2003.

RINEHART, R. E. **Radar for Meteorologists**. Nevada: Rinehart Publishing, 2004.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, v. 20, p. 53–65, 1987.

SHEATHER, S. J. Density estimation. **Statistical Science**, v. 19, n. 4, p. 588–597, 2004.

SHIGA, A. A. **Avaliação de Custos Decorrentes de Descargas Atmosféricas em Sistemas de Distribuição de Energia**. Dissertação (Mestrado) — Universidade de São Paulo, 2007.

SILVA, D. T. A. da; SANTOS, V. F. dos. **ClusteringTools: Uma Ferramenta de Auxílio ao Ensino de Técnicas de Clusterização**. Trabalho de Conclusão de curso (Bacharelado em Ciência da Computação) — Instituto de Matemática da Universidade Federal do Rio de Janeiro, 2007.

SILVA, N. P. da *et al.* AvaliaÇÃo do impacto das descargas atmosfÉricas na qualidade de energia fornecida pelas concessionÁrias: Estudo de caso em uma empresa de distribuiÇÃo de energia do sul do paÍs. **Nucleus**, v. 7, n. 1, p. 139–154, 2010.

SILVA NETO, A. **Tensões Induzidas por Descargas Atmosféricas em Redes de Distribuição de Baixa Tensão**. Dissertação (Mestrado) — Escola Politécnica da Universidade de São Paulo, 2004.

SIQUEIRA, P. H. **Uma Nova Abordagem na Resolução do Problema do Caixeiro Viajante**. Tese (Doutorado) — Universidade Federal do Paraná, 2005.

SOULA, S. Lightning and precipitation. In: BETZ, H. D.; SCHUMANN, U.; LAROCHE, P. (Ed.). **Lightning: Principles, Instruments and Applications**. Netherlands: Springer, 2009.

STEINACKER, R. *et al.* Automatic tracking of convective cells and cell complexes from lightning and radar data. **Meteorology and Atmospheric Physics**, v. 72, p. 101–110, 2000.

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.

UMAN, M. A. **The Lightning Discharge**. New York: Courier Corporation,Dover Publications, 2001.

VIJAYARAGHAVAN, G.; BROWN, M.; BARNES, M. **Practical Grounding, Bonding, Shielding and Surge Protection**. Burlington: Newnes, 2004.

VISACRO, S.; DIAS, R. N.; MESQUITA, C. R. Novel approach for determining spots od critical lightning performance along transmission lines. **IEEE Transactions on Power Delivery**, v. 20, n. 2, p. 1459–1464, 2005.

WESTINGHOUSE, C. S. E. **Electrical Transmission and Distribution Reference Book**. Pennsylvania: Westinghouse Electric Corporation, 1964.

WOOLFORD, D. G.; BRAUN, W. J. Convergent data sharpening for the identification and tracking of spatial temporal centers of lightning activty. **Environmetrics**, v. 18, p. 461–479, 2006.

ZAGOURAS, A. *et al.* Determination of measuring sites for solar irradiance, based on cluster analysis of satellite-derived cloud estimations. **Solar Energy**, v. 97, p. 1–11, 2013.

APÊNDICE A

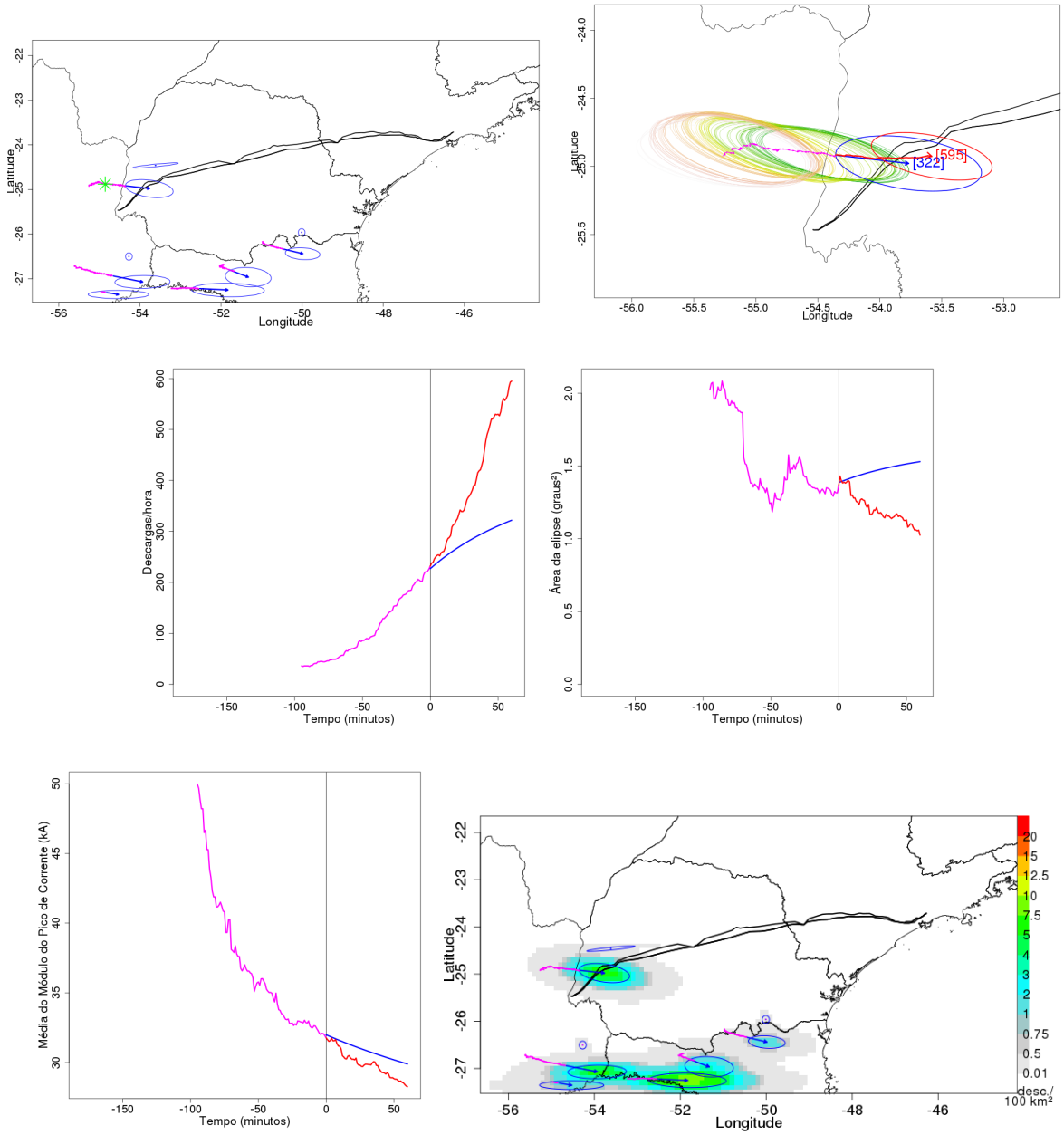


FIGURA 30: 9 tempestades elétricas identificadas em 21/05/2001 às 22:09 ($t = 0$). Houve uma falha no trecho entre Foz do Iguaçu e Ivaiporã às 22:39. Previsão para às 23:09 das variáveis meteorológicas da provável tempestade causadora da falha

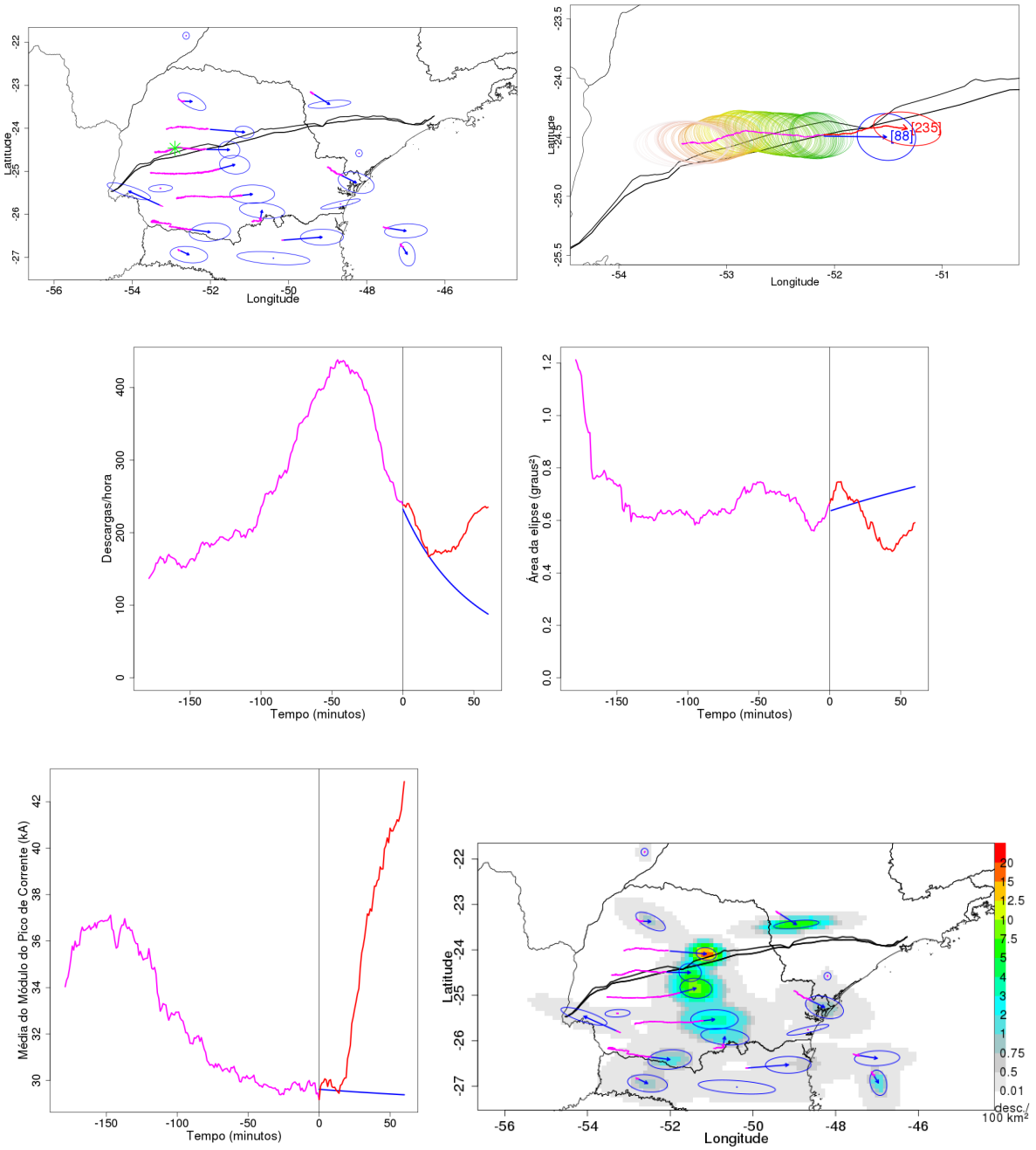


FIGURA 31: 19 tempestades elétricas identificadas em 04/09/2005 às 15:00 ($t = 0$). Houve uma falha no trecho entre Ivaiporã e Itaberá às 15:30. Previsão para às 16:00 das variáveis meteorológicas de uma tempestade candidata à causadora da falha

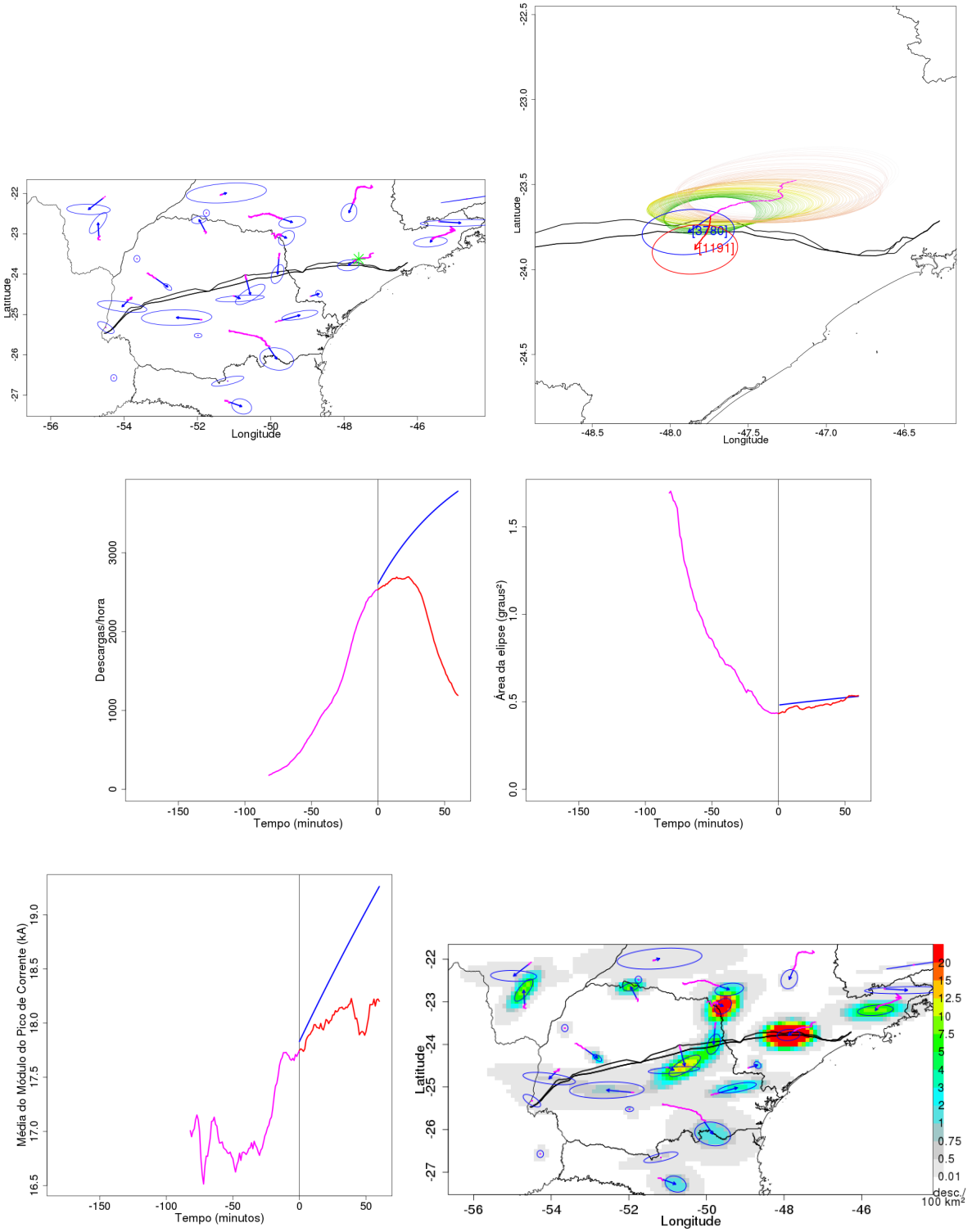


FIGURA 32: 27 tempestades elétricas identificadas em 12/01/14 às 19:00 ($t = 0$). Previsão para às 20:00 das variáveis meteorológicas de uma tempestade selecionada

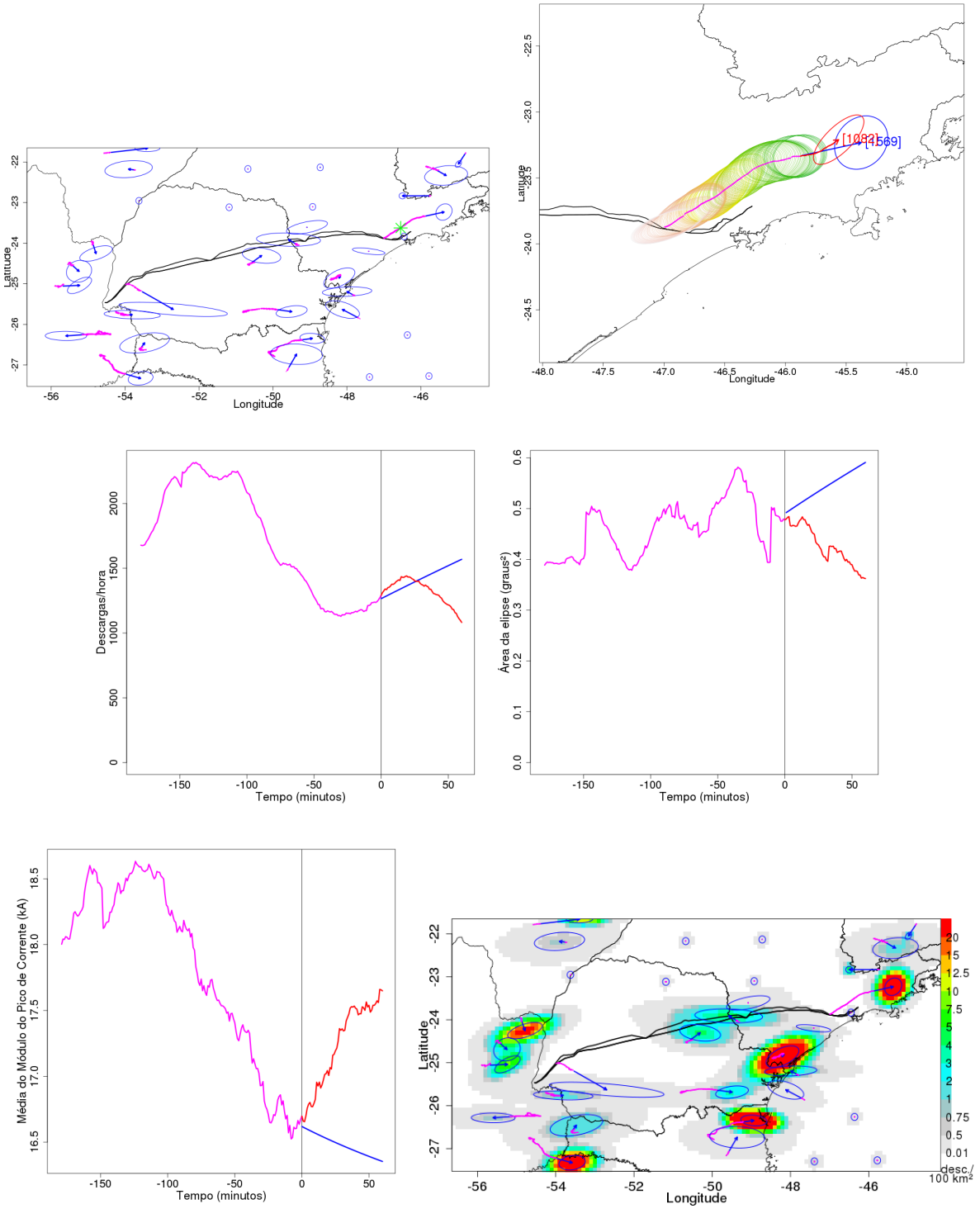


FIGURA 33: 33 tempestades elétricas identificadas em 12/01/15 às 19:00 ($t = 0$). Previsão para às 20:00 das variáveis meteorológicas de uma tempestade selecionada

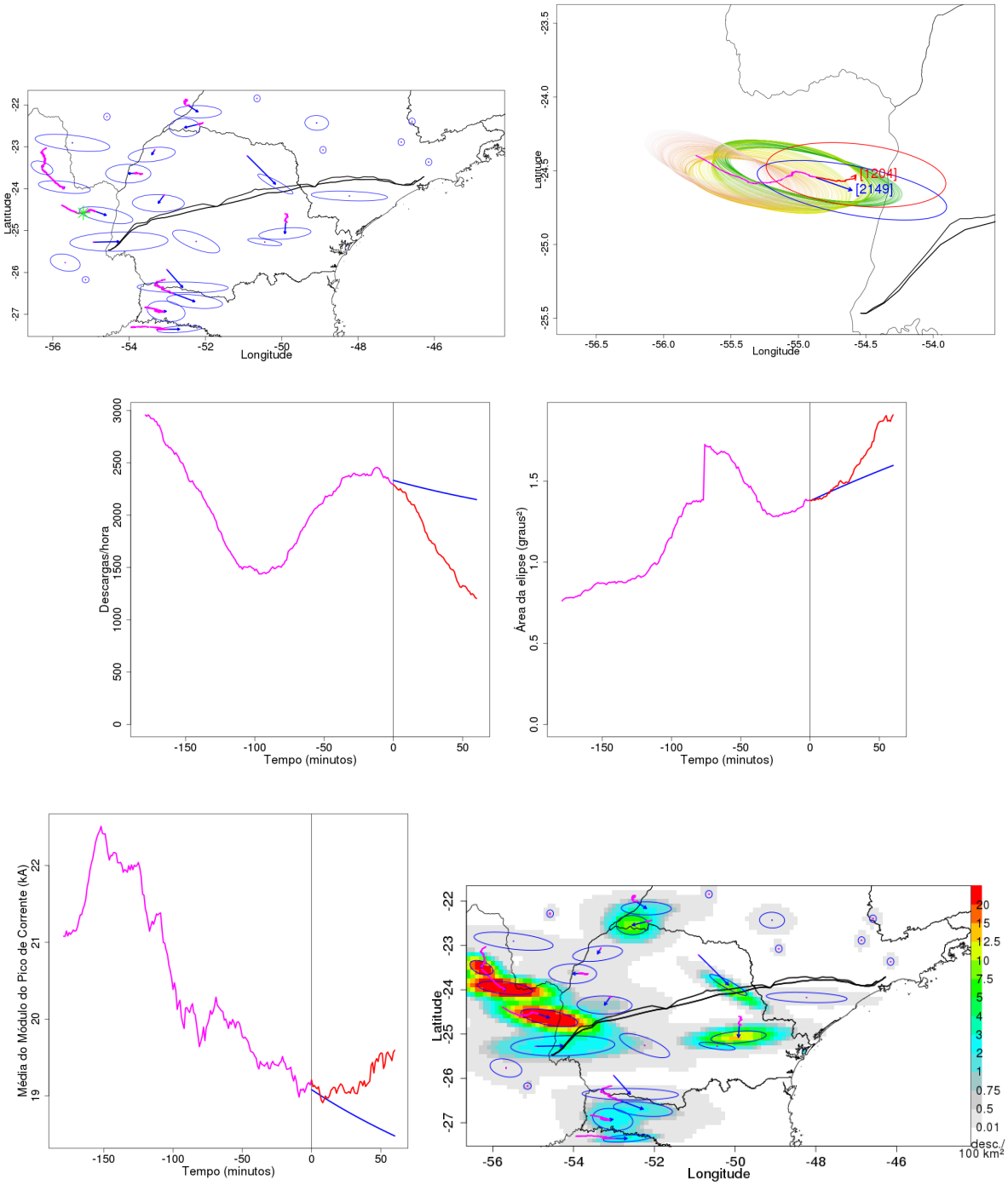


FIGURA 34: 29 tempestades elétricas identificadas em 12/07/15 às 03:00 ($t = 0$). Previsão para às 04:00 das variáveis meteorológicas de uma tempestade selecionada