Análise espaço-temporal de *data streams* multidimensionais

Santiago Augusto Nunes

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura:_____

Santiago Augusto Nunes

Análise espaço-temporal de *data streams* multidimensionais

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências - Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientadora: Profa. Dra. Elaine Parros Machado de Sousa

USP – São Carlos
Julho de 2015

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura:_____

Santiago Augusto Nunes

Spatio-temporal analysis in multidimensional data streams

Master dissertation submitted to the Instituto de Ciências Matemáticas e de Computação - ICMC-USP, in partial fulfillment of the requirements for the degree of the Master Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Elaine Parros Machado de Sousa

USP – São Carlos
July 2015

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

Nunes, Santiago Augusto
N972a Análise espaço-temporal de data streams multidimensionais / Santiago Augusto Nunes; orientadora Elaine Parros Machado de Sousa. -- São Carlos, 2015.
66 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2015.

1. Mineração de Dados. 2. Análise espaço-temporal. 3. Teoria dos Fractais. 4. Data Streams multidimensionais. I. Sousa, Elaine Parros Machado de, orient. II. Título.

Agradecimentos

À minha família que sempre me apoiou em minhas decisões.

À minha orientadora Profa. Dra. Elaine Parros Machado de Sousa por todo o conhecimento transmitido, compreensão e acima de tudo, paciência.

À Dra. Luciana Alvim Romani por todos os conselhos e auxílios, assim como a todos do CEPAGRI-EMBRAPA em Campinas.

Aos amigos do GBDI pelas discussões e conselhos: Willian Denner, Daniel Chino, Lúcio Dutra, Monica, Christian, Marcos Bêdo, Sérgio, Luiz, Bruno, Alceu e todos os demais integrantes.

Aos meus amigos que conviveram comigo e me apoiaram.

À Mayara Fior por todo o carinho e apoio.

Aos funcionários e professores do ICMC-USP.

À FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) pelo apoio financeiro à realização desse trabalho.

Resumo

Fluxos de dados são usualmente caracterizados por grandes quantidades de dados gerados continuamente em processos síncronos ou assíncronos potencialmente infinitos, em aplicações como: sistemas meteorológicos, processos industriais, tráfego de veículos, transações financeiras, redes de sensores, entre outras. Além disso, o comportamento dos dados tende a sofrer alterações significativas ao longo do tempo, definindo *data streams* evolutivos. Estas alterações podem significar eventos temporários (como anomalias ou eventos extremos) ou mudanças relevantes no processo de geração da *stream* (que resultam em alterações na distribuição dos dados). Além disso, esses conjuntos de dados podem possuir características espaciais, como a localização geográfica de sensores, que podem ser úteis no processo de análise. A detecção dessas variações de comportamento que considere os aspectos da evolução temporal, assim como as características espaciais dos dados, é relevante em alguns tipos de aplicação, como o monitoramento de eventos climáticos extremos em pesquisas na área de Agrometeorologia. Nesse contexto, esse projeto de mestrado propõe uma técnica para auxiliar a análise espaço-temporal em *data streams* multidimensionais que contenham informações espaciais e não espaciais. A abordagem adotada é baseada em conceitos da Teoria de Fractais, utilizados para análise de comportamento temporal, assim como técnicas para manipulação de *data streams* e estruturas de dados hierárquicas, visando permitir uma análise que leve em consideração os aspectos espaciais e não espacias simultaneamente. A técnica desenvolvida foi aplicada a dados agrometeorológicos, visando identificar comportamentos distintos considerando diferentes sub-regiões definidas pelas características espaciais dos dados. Portanto, os resultados deste trabalho incluem contribuições para a área de mineração de dados e de apoio a pesquisas em Agrometeorologia.

Palavras-Chave. mineração de dados, análise espaço-temporal, teoria dos fractais, *data streams* multidimensionais.

Abstract

Data streams are usually characterized by large amounts of data generated continuously in synchronous or asynchronous potentially infinite processes, in applications such as: meteorological systems, industrial processes, vehicle traffic, financial transactions, sensor networks, among others. In addition, the behavior of the data tends to change significantly over time, defining evolutionary data streams. These changes may mean temporary events (such as anomalies or extreme events) or relevant changes in the process of generating the stream (that result in changes in the distribution of the data). Furthermore, these data sets can have spatial characteristics such as geographic location of sensors, which can be useful in the analysis process. The detection of these behavioral changes considering aspects of evolution, as well as the spatial characteristics of the data, is relevant for some types of applications, such as monitoring of extreme weather events in Agrometeorology researches. In this context, this project proposes a technique to help spatio-temporal analysis in multidimensional data streams containing spatial and non-spatial information. The adopted approach is based on concepts of the Fractal Theory, used for temporal behavior analysis, as well as techniques for data streams handling also hierarchical data structures, allowing analysis tasks that take into account the spatial and non-spatial aspects simultaneously. The developed technique has been applied to agro-meteorological data to identify different behaviors considering different sub-regions defined by the spatial characteristics of the data. Therefore, results from this work include contribution to data mining area and support research in Agrometeorology.

Keywords. data mining, spatio-temporal analysis, fractals, multidimensional data streams.

Sumário

Lista de Figuras **vi**

Lista de Tabelas **viii**

Lista de Algoritmos **ix**

1 Introdução **1**
1.1 Objetivos . . . 3
1.2 Contribuições . . . 4
1.3 Organização da Dissertação . . . 4

2 Data Streams **5**
2.1 Conceitos Básicos . . . 5
2.2 Mineração de *Data Streams* . . . 7
2.3 Análise espaço-temporal . . . 9

3 Fractais e *Data Streams* **13**
3.1 Fractais . . . 13
3.2 *Stream Instrinsic Dimension Meter* . . . 18

4 *Spatio-Temporal Behaviour Meter* **23**
4.1 Técnicas fundamentais utilizadas pelo método *STB-meter* . . . 24
4.2 Implementação Inicial . . . 27
4.3 Implementação no contexto da ferramenta ClimFractal Analyzer . . . 40
4.4 Considerações Finais . . . 42

5 Experimentos **45**
5.1 Experimentos com Dados Sintéticos . . . 46
5.2 Experimentos com Dados Reais do Estado de São Paulo . . . 48
5.3 Experimentos com Dados Simulados do Estado de São Paulo . . . 53
5.4 Conclusões . . . 55

6 Conclusões **57**
6.1 Contribuições . . . 57
6.2 Publicações . . . 58
6.3 Considerações Finais e Trabalhos Futuros . . . 59

Referências Bibliográficas **61**

Lista de Figuras

3.1 Triângulo de *Sierpinski*. 14
3.2 Reticulado de lados distintos dividindo um Triângulo de *Sierpinski*. 15
3.3 Exemplo da construção de uma *Box Counting Tree* construída sobre um conjunto de dados bidimensional. Traduzida de [Sousa 07b] 15
3.4 Cálculo da dimensão fractal de correlações do triângulo de *Sierpinski* utilizando o método *box-count plot*. 16
3.5 Um segmento de linha imerso em um espaço 3-dimensional tem $D_2 = 1$. 16
3.6 Cálculo da Dimensão Fractal em uma *data stream* multidimensional. 18
3.7 Janela deslizante construída sobre 3 atributos contendo 5 períodos de contagem. 19
3.8 Movimentos realizados sobre a estrutura *Box-Counting Tree*: (a) Ocupação de regiões; (b) Liberação de regiões; (c) Expansão do espaço de representação; (d) Contração do espaço de representação [Sousa 06a]. 21

4.1 Representação da Janela Deslizante estendida nos nós que compõe a *Box-counting tree*. . . . 24
4.2 Representação da Janela Deslizante estendida sobre um *data stream* bidimensional. 25
4.3 Períodos de contagem utilizados na contagem de eventos considerando os últimos 180 dias. . 25
4.4 Períodos de contagem utilizados na contagem de eventos considerando os últimos 180 dias. . 25
4.5 Estrutura de dados hierárquica. 26
4.6 Estrutura da *STB-tree* : Integração de estrutura de dados hierárquica com a árvore de contagem estendida. 26
4.7 Espaço de representação para os atributos espaciais (esquerda) e não espaciais (direita). . . . 31
4.8 *STB-tree* após o processamento do evento $e_1 = \{0,0,4,8\}$. 32
4.9 *STB-tree* após o processamento do evento $e_2 = \{4,4,5,8\}$. 33
4.10 *STB-tree* após o processamento de todos os eventos da Tabela 4.1. 34
4.11 Exemplo de granularidades possíveis para o cálculo de eventos incidentes em uma célula da *STB-tree*. 34
4.12 Espaços de representação após o processamento de todos os eventos da Tabela 4.1. 35
4.13 Evento e_6 que incide fora do espaço de representação da *STB-tree*. 36
4.14 Processo de expansão de um nó da *quadtree* utilizado pela *STB-tree*. 36
4.15 Expansão do espaço de representação da *STB-tree* no sentido do novo evento e_6. 37
4.16 Processo iterativo de expansão do espaço de representação. 38
4.17 *STB-tree* resultante do processo de expansão. 39
4.18 Processo de contração do espaço de representação. 39
4.19 Estrutura *STB-tree* resultante do processo de contração. 40
4.20 Tela principal da Ferramenta *ClimFractal Analyzer* exibindo uma configuração de um *workflow*. 41

4.21 Exemplo de *worlkflow* que realiza o filtro de estações meteorológicas no estado do Rio Grande do Sul. 41
4.22 Exemplo de *worlkflow* que realiza o filtro de estações meteorológicas no estado do Rio Grande do Sul, porém, utilizando uma *quadtree* para realizar a representação da informação espacial. 42

5.1 Correlação das sub-regiões do conjunto de dados sintético. 46
5.2 Correlação das sub-regiões da Região 01. 47
5.3 Dimensão Fractal da regiões do conjunto de dados sintéticos com $n_c = 10$ e $n_e = 2400$. . . . 47
5.4 Dimensão Fractal da sub-região 01. 48
5.5 Disposição das estações meteorológicas disponibilizadas pelo Agritempo. 49
5.6 Filtro utilizado sobre as estações meteorológicas do estado de São Paulo. 50
5.7 Estações meteorológicas resultantes do filtro apresentado na Figura 5.6. 50
5.8 Estrutura hierárquica construída a partir da informação espacial do conjunto de dados. 51
5.9 Cálculo da Dimensão Fractal D_2 utilizando 3 meses de granularidade temporal. 52
5.10 Destaque da Figura 5.9 para sub-região Noroeste do estado de São Paulo. 52
5.11 Destaque da Figura 5.9 para sub-região Sudeste do estado de São Paulo. 53
5.12 Destaque da Figura 5.9 para sub-região Sudoeste do estado de São Paulo. 53
5.13 Comparação da correlação entre dados reais e dados simulados para a região Noroeste do Estado de São Paulo 54
5.14 Comparação da correlação entre dados reais e dados simulados para a região Sudeste do Estado de São Paulo 55
5.15 Comparação da correlação entre dados reais e dados simulados para a região Sudoeste do Estado de São Paulo 56

Lista de Tabelas

4.1 Exemplo de sequência de eventos com compõem um *data stream* multidimensional. 31

5.1 Data Streams. 46
5.2 Descrição dos atributos que compõem o *data stream*. 49

Lista de Algoritmos

4.1 Construção da *STB-tree* 29
4.2 Procedimento *RealizaContagem* 30
4.3 Procedimento *IncrementaContagem* 30
4.4 Procedimento *CalculaContagem* 35
4.5 Procedimento *Expande* 37
4.6 Procedimento *Contrai* 38

Lista de Símbolos

a_s	Conjunto de atributos referentes a informação espacial
a_v	Conjunto de atributos referentes a informação não espacial
$C[]$	Vetor que realiza a contagem de eventos incidentes
D_2	Dimensão Fractal de Correlação
E	Dimensionalidade do Conjunto de Dados
g	Granularidade utilizada para realizar o processamento de contagem na janela deslizante estendida
i_{nc}	Período de contagem atual mantido pela estrutura *STB-tree*
k	Número de níveis
n_c	Número de períodos de contagem
n_e	Número de eventos por período de contagem
N_S	Número de atributos referentes a informação espacial
n_t	Número Total de períodos de contagem da janela deslizante estendida
N_V	Número de atributos referentes a informação não espacial
r_c	Multiplicador que define o intervalo de valores de n_c possíveis na janela deslizante estendida
r_e	Multiplicador que define o intervalo de valores de n_e possíveis na janela deslizante estendida
r_s	Tamanho do lado da *Quadtree* utilizado na representação da espacial
r_v	Tamanho do lado do hiperreticulado utilizado na representação da *Box-counting tree*
S	Estrutura de dados hierárquica
V	Hiperreticulado que representa a *Box-counting tree*

Capítulo 1

Introdução

Tecnologias desenvolvidas na área de Banco de Dados têm sido utilizadas em aplicações que requerem armazenamento persistente, suporte a consultas complexas, atividades de descoberta de conhecimento e de mineração de dados, entre outras tarefas de acesso e análise dos dados. Tradicionalmente, uma base de dados pode ser considerada um conjunto finito e pouco dinâmico de objetos não ordenados, isto é, alterações no estado da base são menos frequentes se comparadas à execução das demais operações mencionadas.

No entanto, esse modelo tradicional de base de dados finita e pouco dinâmica não atende às necessidades de uma classe de aplicações em que a informação é representada como um fluxo contínuo e potencialmente ilimitado de valores, denominado *data stream*. Exemplos de áreas de aplicação que envolvem manipulação e análise de *data streams* incluem: redes de telecomunicação e de dados, tráfego na Internet, *Web logs*, tráfego de veículos, transações financeiras, mercados de ações, sistemas de segurança, redes de sensores, processos industriais, sistemas meteorológicos, aplicações científicas, entre outras.

Em geral, *data streams* são caracterizados por grandes quantidades de dados gerados em processos síncronos ou assíncronos potencialmente infinitos. Por exemplo, um *stream* de valores estimados de pesos de veículos medidos por sensores em uma rodovia de movimento diurno intenso e pouco movimento noturno é gerado por um processo assíncrono; numa outra área de aplicação, um *stream* de medidas de temperatura coletadas em estações meteorológicas de superfície a cada hora é gerado por processo síncrono. Além disso, um *data stream* pode ser unidimensional, com elementos de dados definidos por um único atributo (ex: IDs de pacotes em monitoramento de tráfego de Internet), ou multidimensional, com elementos definidos por um conjunto de atributos (ex: medidas de temperatura, precipitação e pressão atmosférica coletadas em uma estação meteorológica a cada hora).

Outra característica bastante comum em aplicações que envolvem *data streams* é a evolução temporal, ou seja, o comportamento dos dados sofre alterações significativas ao longo do tempo. Essas alterações podem significar eventos temporários (ex: uma onda de calor durante o outono) ou mudanças relevantes no processo de geração do *stream*, podendo resultar em alterações na distribuição dos dados (ex: mudanças climáticas causadas por aquecimento global). A identificação dessas variações de comportamento em *data streams* evolutivos é fundamental em alguns tipos de aplicações, como no

monitoramento de eventos climáticos, onde eventos anômalos e/ou extremos podem significar anomalias como uma onda de calor, ou então um período de *el niño*. Também pode-se citar o monitoramento de processos industriais, onde as variações de comportamento dos dados podem indicar erros ou falhas em alguma parte do processo, e sistemas de cartão de crédito, em que a detecção de variações de comportamento dos dados pode auxiliar detecção de fraude.

Logo, técnicas para análise e mineração de *data streams* evolutivos requerem mecanismos de análise de comportamento que permitam identificar mudanças ao longo do tempo e, consequentemente, permitam atualizar os modelos de conhecimento extraídos dos dados. Entre outras tarefas, a detecção de mudanças de comportamento do *stream*, em particular aquelas associadas a eventos temporários, pode ser utilizada para identificação de anomalias e eventos extremos.

Entretanto, a maior parte das técnicas existentes para a análise de conjuntos de dados tradicionais não se comportam adequadamente em conjuntos de dados com as características presentes em *data streams*, como a possibilidade dos dados analisados serem gerados infinitamente, ou então a impossibilidade de, em muitos casos, realizar mais de uma leitura sobre cada elemento que compõe o *data stream*.

A mineração de *data streams* vem sendo utilizada em diversos domínios de aplicação, que vão de aplicações científicas críticas a aplicações financeiras ou de negócios [Gaber 05]. Em particular, os avanços na tecnologia de sensores, tanto em hardware quanto em software, e a consequente geração de grandes volumes de dados, têm motivado o desenvolvimento de novas técnicas de mineração de *data streams*. A descoberta de conhecimento novo e de interesse a partir de dados de sensores, por exemplo, é relevante para apoiar pesquisas científicas em áreas como a Agrometeorologia, que envolve estudos sobre clima, mudanças climáticas, desastres naturais e o consequente impacto econômico e social [Ganguly 07].

Nesse contexto, em trabalhos anteriores [Nunes 11b, Nunes 11a] foi proposta uma abordagem para análise temporal de dados climáticos baseado em conceitos da Teoria dos Fractais, utilizada como ferramenta de análise de comportamento de *data streams* evolutivos [Sousa 07a]. Técnicas de análise de conjuntos de dados finitos baseadas na Teoria dos Fractais são utilizadas em diversos trabalhos, com bons resultados, como ferramentas de apoio a atividades nas áreas de banco de dados e descoberta de conhecimento, como sintetizado na Seção 3.1. Em particular, o trabalho de Faloutsos e Kamel [Faloutsos 94] mostra que, em geral, dados reais comportam-se como fractais, violando portanto as conjeturas de uniformidade e independência entre atributos assumidas por alguns algoritmos de análise de dados.

De maneira similar, conceitos da Teoria dos Fractais podem ser explorados em algoritmos de análise e processamento de *data streams* multidimensionais [Sousa 07a]. Resultados de trabalhos anteriores mostram que a análise fractal temporal, ou seja, a análise da dimensão fractal medida continuamente ao longo do tempo, é uma ferramenta útil na identificação de mudanças de comportamento dos dados [Sousa 06b, Romani 09]. Mais especificamente, alterações significativas no valor da dimensão fractal indicam variações nas correlações entre os atributos que definem cada elemento do *stream* e, portanto, indicam mudanças de comportamento dos dados.

Essa abordagem é utilizado em [Nunes 11b, Nunes 11a] para identificar eventos extremos, como *El Niño* e *La Niña*, em dados climáticos reais. No entanto, esses trabalhos consideram apenas as medidas das variáveis climáticas e o aspecto temporal dos dados. Em outras palavras, nesses trabalhos as regiões de interesse, definidas pela localização geográfica dos sensores, são pré-selecionadas, e a análise é realizada

considerando somente a evolução temporal dos dados, relacionada às mudanças nas correlações entre as variáveis climáticas, e não a informações espaciais que possam estar associadas aos dados do *stream*.

Técnicas que permitam realizar análises considerando aspectos espaciais e temporais do conjunto de *data streams* são relevantes para análise de dados agrometeorológicos, uma vez que esses dados naturalmente possuem propriedades temporais e espaciais (latitude e longitude, além das variáveis climatológicas, por exemplo), e os estudos dos especialistas da área normalmente envolvem tanto análises mais globais sobre regiões que definem países ou continentes inteiros, quanto análises por sub-regiões, considerando estados por exemplo em diferentes períodos de tempo. Diante desse cenário, este trabalho propõe um método para análise de *data streams* multidimensionais evolutivos que estende a análise fractal temporal para suportar aspectos espaciais dos dados, com aplicação inicial em Agrometeorologia.

1.1 Objetivos

Os objetivos deste trabalho são:

1. Desenvolver um método de análise para *data streams* multidimensionais evolutivos considerando aspectos espaciais e não espaciais dos dados.

2. Aplicar este método a dados agrometeorológicos proveniente de estações meteorológicas de superfície

3. Aplicar este método a dados proveniente de modelo climático a fim de realizar uma comparação entre os resultados obtidos de dados reais.

Para tanto, o trabalho abordará alguns pontos relevantes. São eles:

1. Técnicas de manipulação de *data streams* multidimensionais.

2. A Teoria dos Fractais, mais especificamente, a dimensão fractal, utilizada para realizar a estimativa da correlação dos atributos que compõem o conjunto de dados.

3. Estrutura de dados hierárquicas, utilizadas na representação das informações espaciais contidas nos dados.

Portanto, uma abordagem baseada na Teoria do fractais foi utilizada de modo a detectar mudanças na correlação dos atributos que compõem o *data stream* ao longo do tempo. Mais especificamente, a dimensão fractal de correlação é utilizada para estimar a correlação entre os atributos que compõem o *data stream* multidimensional.

Outro aspecto fundamental é a análise espacial do *data stream*, ou seja, a análise da variação da correlação dos atributos que compõem o conjunto de dados em diferentes regiões do espaço definido por atributos referentes a informações espaciais que façam parte do *data stream*. Para tanto, utilizou-se uma abordagem baseada em estruturas de dados hierárquicas para representar as informações espaciais contidas no conjunto de dados analisados, assim como a *Box Counting Tree* para realizar a representação da informação não espacial.

1.2 Contribuições

Neste trabalho de mestrado, foi desenvolvido um método denominado *Spatio Temporal Behavior meter* (STB-*meter*) que tem como objetivo possibilitar a análise espaço-temporal sobre *data streams* multidimensionais compostos por atributos espaciais e não espaciais. Esse método permite que o especialista do domínio da aplicação detecte mudanças de comportamento ao longo do tempo considerando diferentes regiões definidas pelos atributos espaciais dos dados.

Além disso, outro aspecto relevante do método *STB-meter* é permitir que múltiplas granularidades temporais e espaciais sejam analisadas simultaneamente sem a necessidade de realizar novas leituras do conjunto de dados. Esse aspecto é importante devido a característica presente em *data streams* evolutivos, que podem ser potencialmente infinitas, e desse modo, a persistência do conjunto de dados para futuras leituras se torna inviável. Desse modo, análises de diferentes aspectos espaciais e temporais dos dados podem ser realizadas simultaneamente, possibilitando encontrar anomalias e eventos extremos que ocorram em diferentes escalas de tempo e espaço.

O método *STB-meter* foi aplicado, inicialmente, sobre dados sintéticos, de modo a validar o método em conjuntos de dados com um comportamento conhecido, em seguida o método foi aplicado sobre dados meteorológicos provenientes de estações meteorológicas de superfície. As análises dos especialistas dos domínio sugere que o método pode ser utilizado para detectar diferenças entre as correlações dos atributos dos dados em diferentes regiões do espaço, de modo a auxiliar em suas análise e na tarefa de tomada de decisões. Foi possível detectar como eventos extremos como *El Niño* alteram de forma diferente as correlações entre as variáveis climatológicas, tendo assim mais efeitos sobre algumas regiões que em outras. Por fim, o método foi aplicado sobre dados provenientes de modelo climático, a fim de realizar a comparação entre a correlação existente entre as variáveis climatológicas provenientes de estações meteorológicas de superfície em relação a dados proveniente do modelo.

1.3 Organização da Dissertação

Inicialmente, as definições de *data streams* são apresentadas no Capítulo 2, apresentando algumas técnicas convenientes na manipulação desse tipo de dados assim como alguns trabalhos relacionados. No Capítulo 3 são apresentados os principais conceitos de Teoria do Fractais e sua aplicações na análise de correlação de atributos em *Data Streams* multidimensionais.

No Capítulo 4 é apresentado o método *Spatio Temporal Behaviu Meter* proposto neste Trabalho de Mestrado. Esse método tem como objetivo realizar o cálculo da dimensão fractal em diversas granularidades temporais e espaciais em *Data Streams* multidimensionais com atributos espaciais e não espaciais. A seguir, no Capitulo 5, alguns experimentos são apresentados de modo a demonstrar a eficácia do método assim como alguns resultados obtidos e, por fim, no Capítulo 6, são discutidos os resultados obtidos assim como os passos seguintes para trabalhos futuros possíveis sobre o método proposto neste Trabalho de Mestrado.

Capítulo 2

Data Streams

Neste capítulo são apresentados os principais conceitos de *Data Streams* utilizados neste trabalho.

2.1 Conceitos Básicos

Pode-se definir um *data stream*, ou fluxo de dados, como uma sequência contínua de itens de dados ou eventos, ordenada e potencialmente infinita. Mais especificamente, um *data stream* pode ser definido como uma sequência de eventos $< e_1, e_2, ..., e_N, ..., >$ em que cada evento é representado por um conjunto de E valores (atributos), tal que $< e_i = (a_i, ..., a_E) >$. Usualmente, os eventos são gerados e processados em tempo real.

Pode-se dividir *data streams* em duas categorias gerais [Chaudhry 05]:

1. ***Data streams* de transação :** Estes tipos de *data streams* são *logs* resultantes da interação entre entidades de sistema. Como exemplo pode-se citar a interação dos usuários com um *web site*, onde as ações do usuário são registradas em forma de *logs*. Esses *logs* são monitorados por aplicações com diversos objetivos, por exemplo, analisar a performance do servidor, ou então realizar a personalização do *site* de acordo com as preferências do usuário. As informações de interação do usuário com o *site* são continuamente armazenados.

2. ***Data streams* de medidas:** Estes tipos de *data streams* são resultados do monitoramento do estado de entidades de interesse. Com a queda no preço de sensores assim como sua miniaturização, diversas atividades e ambientes passaram a ser monitoradas. O resultado desse monitoramento são *streams* de medidas, como o monitoramento de soldados em um campo de batalha, de carros em rodovias assim como o monitoramento do clima. Como outro exemplo, pode-se citar as redes de comunicação em larga escala, onde ocorre o monitoramento contínuo com vários objetivos, como localizar falhas na comunicação, detectar gargalos de recursos assim como detectar ataques e intrusões na rede.

Em ambos os casos, os eventos que compõem a *data stream* apresentam uma determinada ordem e essa ordem é relevante no momento em que esse *stream* é processado e pode ser influenciada por diversos fatores, como a latência na aquisição dos dados, falha dos sensores, erros de leitura, etc.

Nesse contexto, pode-se distinguir a ordem em *data streams* em duas categorias:

1. **Implícita -** adicionada ao evento quando este é recebido no sistema que processa o *data stream*. Neste caso, a marca de tempo pode ser simplesmente um número sequencial que determina uma ordem total entre os eventos do *stream*.

2. **Explícita -** é uma propriedade do próprio evento, isto é, a marca de tempo é um de seus atributos. *Data streams* definidos por eventos do mundo real tipicamente possuem marcas de tempos explícitas. Como exemplo pode-se citar a data e hora de aquisição de dados climáticos em um sensor meteorológico.

Em diversas aplicações, os eventos que compõem o *data stream* são formados por diversos atributos, por exemplo, *streams* gerados a partir de sensores em uma rodovia, que capturam a velocidade do veículo e sua placa geram um evento $e_t = (a_1, a_2)$ sendo a_1 a velocidade do veículo e a_2 a placa do carro, medidas no momento t. Esse tipo de *data stream* é denominado multidimensional devido ao fato de que cada evento possui sua dimensão $E = 2$. Do mesmo modo, dados meteorológicos apresentam uma série de medições, onde cada uma delas é referente a uma variável meteorológica, como temperatura e precipitação. Portanto, *data streams* meteorológicos usualmente são multidimensionais e além disso, o processo de aquisição de dados pode ser executado indefinidamente, gerando assim um conjunto de dados potencialmente infinito.

Desse modo, uma série de características inerentes a *data streams* devem ser consideradas no processo de análise [Gama 10], sendo elas :

1. O fluxo de dados é contínuo e potencialmente infinito.

2. Os dados são transitivos.

3. Os dados podem evoluir com o passar do tempo.

4. Uma vez recebido um evento do *stream*, o mesmo deve ser descartado ou armazenado, não sendo trivial realizar mais que uma leitura desse dado.

Nesse cenário, sistemas de processamento e análise de *data streams* contínuos deve-se considerar o fato de que estes são potencialmente infinitos. Usualmente, o foco de análise do *data stream* está em uma sequência de eventos, e não no *stream* inteiro, sendo necessário limitar o escopo de análise dos dados.

Nesse contexto, o modelo baseado em janelas permite determinar um escopo sobre a sequência de eventos que compõem o *data stream*. Essas janelas são definidas basicamente em função de dois critérios: tamanho e escopo [Chaudhry 05]:

- Tamanho da janela - a **janela baseada em tempo**, também chamada de janela física, tem seu tamanho definido por um intervalo de tempo. Por exemplo: em um sistema que monitora o tráfego em rodovias, com eventos gerados por medidas de velocidade de cada veículo ao passar por um determinado ponto, uma consulta contínua do tipo “obter a velocidade média dos veículos nos

últimos 10 minutos" requer uma janela baseada em tempo. Desse modo, a consulta é executada sobre os eventos contidos na janela e reavaliada continuamente, sempre considerando os eventos ocorridos nos 10 minutos mais recentes. Uma abordagem alternativa é a **janela baseada em eventos**, também chamada de janela lógica, definida por uma quantidade de eventos pré-determinada de acordo com as particularidades do domínio da aplicação.

- Escopo da janela - a **janela móvel** (*sliding window*) tem tamanho fixo, mas seus pontos de início e fim movimentam-se para substituir eventos antigos à medida que novos eventos são recebidos. Seguindo o exemplo do sistema de monitoramento de tráfico, a janela movimenta-se de maneira a abranger os eventos recebidos nos últimos 10 minutos e descartar os mais antigos. Numa outra abordagem, a **janela por marcação** (*landmark window*) tem um de seus pontos, inicial ou final, mantido fixo, enquanto o outro movimenta-se. No exemplo, para uma consulta do tipo "calcular a velocidade média dos veículos desde a ocorrência de um acidente na rodovia", é mais apropriada a utilização de janelas por marcação.

Outro ponto relevante é que *data streams* podem ser gerados por processos síncronos, por exemplo a aquisição da velocidade de uma conexão na internet realizada a cada intervalo de tempo definido, ou então por processos assíncronos, por exemplo a aquisição do peso de um caminhão em um posto de pesagem, realizado somente quando um caminhão entra no posto.

Esses processos podem ser executados indefinidamente fazendo com que conjuntos de dados gerados por eles sejam muito volumosos e, possivelmente, infinitos. Essa característica traz novos desafios na tarefa de análise e mineração de *data streams*, pois, de maneira diferente dos métodos tradicionais de análise onde o conjunto de dados analisados pode ser considerado estático e está sempre disponível, permitindo realizar diversas leituras sobre o conjunto de dados, em métodos de análise de *data streams*, o conjunto de dados normalmente não pode ser armazenado em memória principal e, além disso, em ambientes que tratam *data stream* transientes, que não são armazenados em disco, os eventos podem ser acessados apenas uma vez e devem ser analisados assim que obtidos e depois descartados.

Desse modo, muitos dos algoritmos desenvolvidos para conjuntos de dados tradicionais não se comportam adequadamente para conjuntos de dados com as característica presentes em *data streams*.

Na Seção 2.2 é apresentada uma breve discussão sobre a tarefa de mineração de *data streams* assim como os novos desafios presentes devido às características desses conjuntos de dados.

2.2 Mineração de *Data Streams*

Nas últimas décadas, pesquisas na área de descoberta de conhecimento têm sido focadas em lotes de dados, usualmente utilizando pequenos conjuntos de dados. Na aprendizagem de máquina sobre lotes de dados, todo o conjunto de dados de treinamento está disponível para os algoritmos, que apresentam resultados de modelos de decisão após processar todo conjunto de dados, eventualmente, múltiplas vezes. A ideia utilizada nessa abordagem é que os conjuntos de dados de treinamento são gerados aleatoriamente de acordo com alguma distribuição estacionária [Gama 10].

Em muitas aplicações, algoritmos de descoberta de conhecimento são necessários em ambientes dinâmicos, onde os dados são coletados ao longo do tempo. Uma propriedade importante é que esses algoritmos considerem e incorporem os novos dados em sua análise. Alguns algoritmos supervisionados

são naturalmente incrementais, como o *k-nearest neighbors* e o *naive-Bayes*. Por outro lado, algoritmos baseados em árvores de decisão requerem mudanças substancias para realizar uma análise incremental. Além disso, se o processo analisado não apresenta um conjunto de dados com distribuição estritamente estacionária, a distribuição dos dados pode mudar ao longo do tempo. Desse modo, considerar e incorporar novos dados é uma propriedade necessária na mineração de *data streams*, porém, não suficiente pois é necessário também que existam mecanismos para incorporar mudanças da distribuição dos dados e também considerar a possibilidade de dados faltantes.

Diferente da abordagem usual, a tarefa de descoberta de conhecimento em *data streams* considera que a distribuição dos dados pode mudar, assim como podem existir dados faltantes. Nesse contexto, diversos algoritmos para a manipulação de *data stream* foram desenvolvidos. Dentre esses, pode-se citar as seguintes tarefas principais [Gama 10]:

- Detecção de Mudanças: Em muitas aplicações, as tarefas de mineração de dados agem em ambientes dinâmicos onde o fluxo de dados é constante. Se esses dados não admitem uma distribuição de probabilidade estritamente estacionária, os conceitos adquiridos na análise inicial dos dados podem mudar ao longo do tempo. No entanto, muitos dos trabalhos de mineração de dados assumem que o conjunto de dados de treinamento é gerado a partir de uma distribuição estacionária de probabilidades. Nesse contexto, detectar mudanças na distribuição de dados pode representar uma informação relevante em várias aplicações, como em sistemas de segurança, telecomunicações, transações bancárias. Em [Kifer 04] é apresentado uma técnica baseada em funções de distâncias entre duas janelas, uma contendo dados provenientes do *data stream* e outra contendo dados de referências, de modo a detectar mudanças na distribuição do *stream*, sendo atualizada sempre que uma mudança é detectada.

- Agrupamentos: A tarefa de Agrupamento em *data streams* pode ser definida como "a manutenção consistente de agrupamentos realizada continuamente sobre sequência de eventos observados por longos períodos, utilizando uma pequena quantidade de memória e tempo". Os desafios na tarefa de agrupamentos sobre *data streams* são referentes ao fluxo de dados contínuo e também a necessidade de analisá-los em tempo real. Essa característica faz com que os algoritmos de agrupamento em *data streams* sejam incrementais, mantendo uma estrutura de agrupamentos que evolui com o tempo. Isso acontece porque *data streams* evoluem com o tempo e grupos podem deixar de existir assim como novos grupos podem surgir, refletindo assim a característica dinâmica de *data streams*. Em [Rodrigues 07] utiliza-se uma abordagem incremental para realizar o agrupamentos de *streams*. Nessa abordagem, uma estrutura baseada em árvores construída de modo a agrupar *sstreams* considerando a similaridade entre eles, onde a similaridade é baseada no coeficiente de correlação de Pearson. Em [Guha 03] é utilizada uma abordagem baseada em *k-median* de modo a detectar grupos no conjunto de dados, além de apresentar uma solução que permite determinar o parâmetro *k* utilizado na tarefa de agrupamento. Em [Aggarwal 04] é proposto um método de para agrupamento de *data streams* multi-dimensionais de alta dimensionalidade.

- Detecção de Padrões Frequentes: Ultimamente, mineração de padrões frequentes é uma das áreas de maior pesquisa dentro da área de descoberta de conhecimento, assim, um grande número de algoritmos foram desenvolvidos. A mineração de padrões frequentes em *data streams* apresenta um novo desafio. Além do fato de algoritmos sobre *data streams* terem que realizar apenas uma

leitura sobre os dados, a memória limitada e a impossibilidade de fazer uma análise combinatória sobre os dados aumenta esse desafio. Na tarefa de mineração de padrões frequentes sobre *data streams*, pode-se identificar três abordagens principais: as técnicas que não distinguem elementos novos dos antigos; as que dão maior importância aos eventos recentes; e as que utilizam diferentes granularidades temporais na tarefa de mineração. Em [Papadimitriou 04] é proposto o método AWSOM, que tem como principal objetivo detectar padrões e tendências de interesse sobre o *stream*.

- Classificação: O problema de classificação é um dos mais amplamente estudados no contexto de mineração em *data stream*. A principal dificuldade é tratar dados que evoluem com o tempo, onde é possível que novas classes apareçam assim como classes antigas deixem de existir. Um levantamento de diversos algoritmos de classificação em *data streams*, seguindo diferentes abordagens, pode ser encontrado em [Aggarwal 07, Aggarwal 14].

- Detecção de Padrões Novos: Detecção de padrões novos é uma tarefa de mineração de dados que consiste em identificar padrões novos ou desconhecidos que o sistema de aprendizagem não detectou durante a fase de treinamento. Essa tarefa é fundamental para um bom sistema de agrupamentos ou de classificação, pois muitas vezes os conjuntos de dados utilizados na fase de treinamento não contêm o conhecimento de todos os padrões possíveis existentes em uma aplicação.

- Mineração de Regras de Associação: A tarefa de mineração de regras de associação em *data streams* apresenta novos desafios. Primeiramente, não é possível realizar múltiplas leituras do conjunto de dados, tendo em vista que *data streams* podem ser potencialmente infinitos. Desse modo, também não é possível armazenar o *data stream* em memória. Outro desafio é decorrência de que a distribuição de um *data stream* pode mudar com o tempo. Essa característica faz com que os métodos para mineração de regras de associação tenham que se adaptar para alteração na distribuição do conjunto de dados [Jiang 06]. Em [Kosina 15] é apresentado um algoritmo para extração de regras de associação sobre *data streams* que utiliza um classificador reprojetado de modo que seja capaz de extrais regras de associação sobre cada classe.

- Detecção de Anomalias: A tarefa de detecção de anomalias em *data streams* apresenta os mesmos desafios presentes na tarefa de mineração de regras de associação. Além disso, o *stream* normalmente contém mais eventos normais do que anômalos, o que faz com que raramente exista um conjunto de dados de treinamento com todos os casos anômalos possíveis [Tan 11].

Na mineração de dados em *data streams*, pode-se citar ainda trabalhos que incluem análises de tendências em *data streams* evolutivas [Aggarwal 03], onde é apresentado um *framework* para detectar tendencias espaciais e temporais sobre *data streams* multidimensionais.

2.3 Análise espaço-temporal

Na tarefa de mineração de *data streams*, a abordagem baseada em janelas permite limitar o escopo da análise sobre uma *stream* a uma sequência de eventos, desse modo, com o deslocamento da janela, novos eventos passam a ser considerados enquanto eventos mais antigos são descartados. Como resultado dessa

abordagem, uma análise temporal pode ser realizada sobre a *data stream*, ou seja, uma análise é realizada sobre uma sequência de eventos que é atualizada ao longo do tempo. Por exemplo, em um *data stream* gerada por um sensor que realiza a aquisição da velocidade de carros em uma rodovia, uma análise que realiza o cálculo da velocidade média a cada minuto utiliza uma janela deslizante para limitar o escopo de análise, a cada minuto um novo cálculo da média de velocidade é realizado e, deste modo, a evolução da velocidade média ao longo do tempo é obtida.

A análise temporal realizada por meio da utilização de janelas deslizantes pode ser estendida de modo a considerar outros aspectos do *data stream*, no exemplo citado anteriormente, a análise realizada sobre o *stream* gerada por um sensor localizado em uma rodovia leva em consideração apenas as velocidades obtidas pelo sensor, porém, outras informações intrínsecas ao sensor podem ser utilizadas de forma conjunta. Esse sensor está localizado em um trecho específico da rodovia e desse modo, possui uma informação espacial relacionada a ele.

Nesse cenário, dado um conjunto de sensores localizados em uma rodovia, é possível utilizar a informação espacial de cada sensor juntamente com as aquisições das velocidades para expandir a análise temporal de modo a considerar também aspectos espaciais do conjunto de dados. Considerando essas novas informações, é possível analisar, por exemplo, como a variação da velocidade média ocorre em diferentes localidades.

Este novo conjunto de informações permite que o especialista do domínio melhore sua análise sobre o conjunto de dados analisados, podendo, neste cenário, utilizar essa análise espaço-temporal para detectar problemas que ocorrem entre as localidade analisadas, por meio da avaliação continuada da velocidade média em diferentes locais. Esse tipo de análise é considerado um novo desafio na área de descoberta de conhecimento [Gama 10] e, desse modo, diversas pesquisas tem como foco a análise espaço-temporal em *data streams* multidimensionais

Na Seção 5 são apresentados uma série de experimentos sobre dados meteorológicos reais e também dados provenientes de modelos climáticos. Uma das características principais de dados climáticos é que conjuntos de dados provenientes de fontes próximas, espacialmente e temporalmente, tendem a serem mais similares do que dados que são provenientes de fontes distantes [Faghmous 14]. Essa característica, conhecida como *autocorrelação* faz com que dados com características espaço-temporais que considerem o conjunto de dados estatisticamente independentes ou identicamente distribuídos não sejam adequadas.

Esses conjuntos de dados apresentam novos desafios na área de descoberta de conhecimento, tendo em vista que as áreas de pesquisa dados com informações unicamente espaciais, ou unicamente temporais é amplamente abordada na literatura. A mineração espaço-temporal sobre conjuntos de dados climáticos considerando esses dois aspectos simultaneamente é uma área emergente [Faghmous 14]. Nesse cenário, uma série de novas técnicas e algoritmos devem ser desenvolvidas de modo a permitir a análise desses conjuntos de dados.

Em [Faghmous 12], séries temporais contendo informações referentes a altura da superfície do oceano são utilizadas para detectar a ocorrência de redemoinhos. Nesse trabalho, inicialmente realiza-se a seleção de série temporais que apresentem características de comportamento presentes em redemoinhos. Em uma segunda etapa, utiliza-se a informação espacial referente às séries temporais selecionadas para obter dados referentes as proximidades dessa região. Por fim, as séries temporais são utilizadas para detectar e monitorar o movimento dos redemoinhos, de modo que quando detecta-se que ouve

uma mudança de posicionamento ao longo do tempo, novas séries temporais vizinhas passam a ser consideradas na análise.

Em [Sigrist 12] é utilizado um modelo espaço-temporal estatístico para realizar a previsão de precipitação. Nesse trabalho, foi construído um modelo bayesiano hierárquico para conjunto de dados com características espaço-temporais, que são referentes a precipitação e outras variáveis relacionadas, como por exemplo o movimento de convecção. Esse modelo é utilizado para realizar a previsão com 3 horas de antecedência, baseando-se em dados provenientes de modelos climáticos e também de dados meteorológicos reais.

Na tarefa de detecção de tendências, podemos citar o trabalho de [Meng 06] é utilizado uma abordagem baseada em cadeias de Markov para detectar tendências em *Data Streams* com informações espaço temporal.

Na tarefa de detectar correlações entre sub-regiões do espaço, pode-se citar o trabalho [Kawale 12]. Este trabalho tem como objetivo detectar anomalias na pressão atmosféricas de duas regiões distantes entre si, onde exista uma correlação negativa entre elas. Para tanto, técnicas estatísticas são utilizada de modo a extrair séries temporais representando padrões de *teleconnections*, isto é, anomalias climáticas distantes e interconectadas.

Considerações Finais

Neste Capítulo foram apresentados os principais conceitos relacionados a *Data Streams* e análise espaço-temporal, assim como alguns trabalho relacionados a área. Na Seção 3 são apresentados os principais técnicas utilizadas neste trabalho de mestrado.

Capítulo 3

Fractais e *Data Streams*

Tendo em vista os conceitos de *data streams* apresentados no Capítulo 2, neste capítulo é apresentada uma abordagem para análise da correlação de atributos em *data streams* multidimensionais baseada na Teoria dos fractais.

A Teoria dos fractais oferece uma série de ferramentas que permitem mensurar o nível de correlação de um conjunto de dados, como é apresentado na Seção 3.1. Neste trabalho, técnicas presentes na Teoria dos Fractais são utilizadas para mensurar a correlação entre atributos de uma *data stream* composta por séries temporais climáticas. Em [Sousa 07a] é proposta uma técnica de cálculo continuado da dimensão fractal em *streams* evolutivas. Esta proposta será utilizada neste trabalho para realizar a análise fractal de séries temporais climáticas. O algoritmo utilizado para realizar o cálculo é o *SID-meter* apresentado com maiores detalhes na Seção 3.2.

A variação da correlação ao longo do tempo pode representar informações importantes de modo a auxiliar o especialista do domínio da aplicação na tarefa de tomada de decisões.

3.1 Fractais

Um fractal é definido pela propriedade de auto-similaridade, isto é, apresenta as mesmas características para diferentes variações em escala e tamanho. Portanto, partes do fractal (seja ele uma estrutura matemática, um objeto ou um conjunto de dados) são similares, exata ou estatisticamente, ao fractal como um todo [Schroeder 91]. A Figura 3.1 ilustra alguns passos do processo de construção do Triângulo de *Sierpinsky*, um fractal matemático clássico. É possível observar que o triângulo resultante possui "buracos" em qualquer escala, e que cada um dos triângulos interiores é uma "miniatura" do triângulo todo.

Os fractais geralmente possuem características incomuns, e até mesmo paradoxais. O Triângulo de *Sierpinsky*, por exemplo, tem perímetro infinito (proporcional a $lim_{i \longrightarrow \infty}(1+1/2)^i$) e área nula (proporcional a $lim_{i \longrightarrow \infty}(3/4)^i$), pois a cada iteração de seu processo de construção, teoricamente infinito, o perímetro aumenta e a área diminui. Com tais propriedades, este fractal não pode ser considerado um

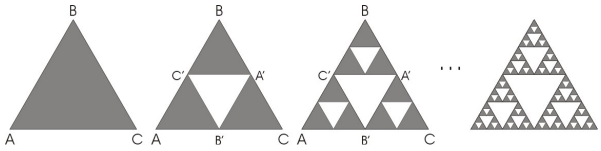

Figura 3.1: Triângulo de *Sierpinski*.

objeto Euclideano unidimensional (caso contrário teria perímetro finito) e nem um objeto Euclideano bi-dimensional (visto que possui área nula).

Nesse sentido, é factível considerar uma dimensionalidade fracionária, chamada de dimensão fractal [Mandelbrot 83]. Na Teoria dos Fractais existem diversas definições para dimensão fractal, a medida de dimensão fractal mais trivial é voltada para objetos exatamente auto-similares, isto é, fractais compostos por M réplicas menores dele mesmo, onde cada réplica segue uma escala de $1 : s$ do fractal inicial.

Dimensão Fractal $\mathcal{D}$:

Seja M o número de réplicas e s o fator em que cada réplica é reduzida a partir do fractal original. A dimensão $\mathcal{D}$ de um fractal exatamente auto similar definido em um espaço E-dimensional é:

$$\mathcal{D} \equiv \frac{\log(M)}{\log(s)} \tag{3.1}$$

Intuitivamente, a dimensão fractal do Triângulo de *Sierpinsky* é um valor entre 1 e 2; matematicamente, 1,58 é o valor exato. Contudo, a dimensão fractal $\mathcal{D}$ pode ser obtida apenas de fractais exatamente auto-similares, tendo assim uma regra de construção bem definida, de modo que seja possível determinar os valores de M e s na Equação 3.1.

Para os conjuntos de fractais estatisticamente auto-similares, cuja não existe uma regra de formação bem definida, é possível obter uma estimativa da dimensão fractal por meio do método *Box-counting*.

Método Box-counting : seja $\mathcal{P}$ um conjunto de pontos imerso em um espaço E-dimensional dividido por um hiper-reticulado com células de lado r, e seja $N(r)$ o número de células que contenham mais que 1 ponto incidentes. A dimensão $\mathcal{D}_0$ para fractais finitos, denomina Dimensão Fractal de Hausdorff, é definida por:

$$\mathcal{D}_0 \equiv -\lim_{r \to 0} \frac{\log(N(r))}{\log(r)} \tag{3.2}$$

Na Figura 3.2 é possível verificar o resultado da diminuição de r sobre o número de células $\log(N(r))$ ocupadas, ou seja, o número de células ocupadas diminui conforme a granularidade se torna mais fina, considerando células com lado r menores.

Na teoria dos fractais existem diversas definições para a dimensão fractal, como a Dimensão de Correlação Fractal D_2, apresentada na Equação 3.3. Essa definição é uma especialização de uma família de dimensões fractais utilizada para conjuntos estatisticamente auto-similares denominada Dimensão Fractal Generaliza D_q.

Dimensão de Correlação Fractal $\boldsymbol{D_2}$: seja $\mathcal{P}$ um conjunto de pontos imerso em um espaço E-dimensional dividido por um hiper-reticulado com células de lado r, e seja $p_{i,r}$ a contagem de pontos que

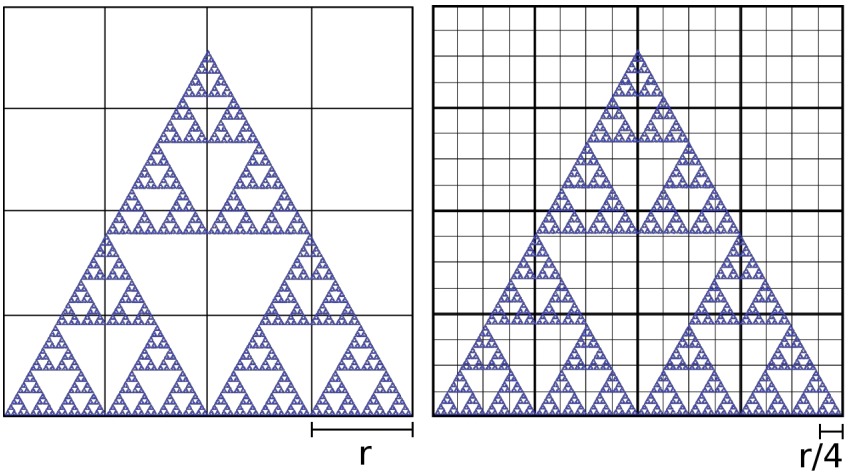

Figura 3.2: Reticulado de lados distintos dividindo um Triângulo de *Sierpinski*.

incidem na célula i. A dimensão fractal D_2 é definida por:

$$D_2 \equiv \frac{\partial \log(\sum_i p_{i,r}^2)}{\partial \log(r)} = constante \qquad r \in (r_1, r_2) \tag{3.3}$$

Naturalmente, o método *box-counting* pode ser estendido de modo a utilizar a dimensão fractal D_2, ou seja, cada célula do hiper-reticulado recebe a contagem de pontos incidentes na $i-$nésima célula ao quadrado, ou seja, $p_{i,r}^2$.

Em [Traina 00] é apresentado o algoritmo **LIBOC()** (*Linear Box Occupancy Counter*) que realiza o cálculo da dimensão fractal D_2 por meio de uma abordagem baseada no método *Box couting* tendo uma complexidade linear *(O(N))* em relação ao número de pontos N. Esse algoritmo é baseado em uma estrutura de hiper-reticulados com dimensão E em múltiplos níveis de refinamento, onde E é o número de atributos do conjunto de dados.

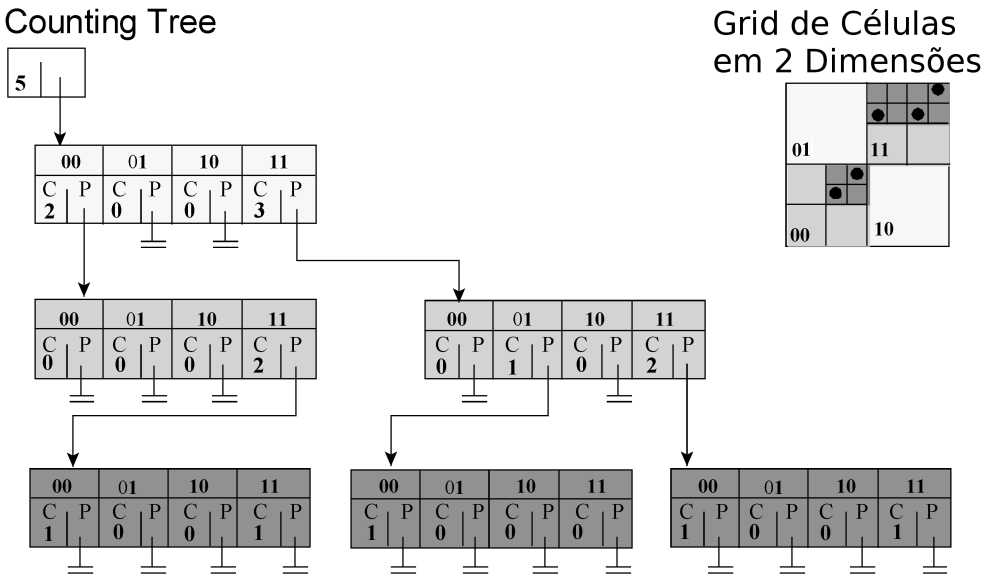

Figura 3.3: Exemplo da construção de uma *Box Counting Tree* construída sobre um conjunto de dados bidimensional. Traduzida de [Sousa 07b]

Na Figura 3.3 é apresentado uma *Box Counting tree*, que é utilizada pelo algoritmo **LIBOC()**, construída sobre um conjunto de dados bi-dimensional. Pode-se verificar que cada nível da estrutura representa um nível de refinamento do hiper-reticulado utilizado no método *Box Counting*. Além disso, cada célula mantém a contagem de pontos incidentes do conjunto.

Uma forma de obter a dimensão fractal é por meio da construção do *box-count plot*, onde é representado em um gráfico os valores de $\log(N(r))$ por $\log(r)$. Como o valor de $\log(N(r))$ diminui com o aumento de $\log(r)$, a curva resultante é decrescente e linear em um intervalo (r_1, r_2). Desse modo, o valor da dimensão fractal é dado pelo negativo do valor da inclinação da reta que melhor se ajusta a curva no intervalo em questão.

Na Figura 3.4 é possível verificar o *box-count plot* construído utilizando o algoritmo **LIBOC()** sobre um conjunto de dados formado pelo triângulo de *Sierpinski*.

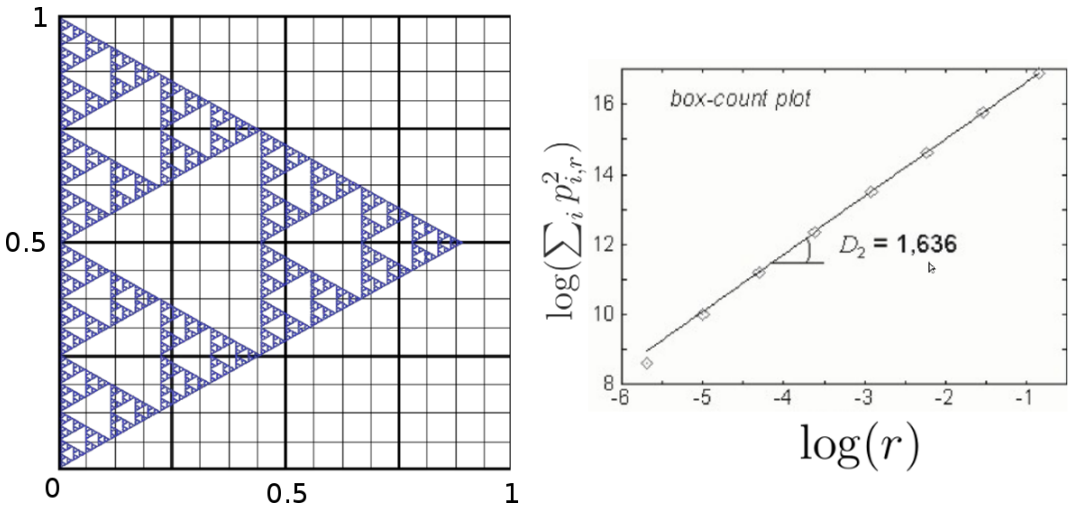

Figura 3.4: Cálculo da dimensão fractal de correlações do triângulo de *Sierpinski* utilizando o método *box-count plot*.

Efetivamente, a dimensão fractal fornece uma estimativa da dimensão intrínseca, que corresponde à dimensionalidade do conjunto de dados, independente da dimensão do espaço em que o conjunto está imerso. Por exemplo: um conjunto de pontos dispostos como uma linha possui dimensão intrínseca igual a 1, mesmo que esteja imerso em um espaço de dimensão maior ou igual a 2. Na Figura 3.5 é apresentado um exemplo onde um segmento de reta, que tem $D_2 = 1$, imerso em um espaço de três dimensões.

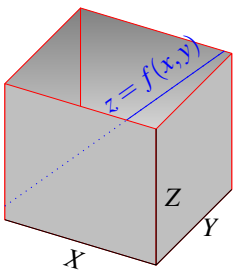

Figura 3.5: Um segmento de linha imerso em um espaço 3-dimensional tem $D_2 = 1$.

Conjuntos de dados reais, em geral, apresentam a propriedade da auto-similaridade, podendo ser

considerados conjuntos fractais [Faloutsos 94]. A vantagem imediata dessa conjetura é que a dimensão intrínseca de um conjunto real finito pode ser mensurada por meio de sua dimensão fractal D_2.

Usualmente, dados reais imersos em espaços de alta dimensionalidade apresentam baixa dimensão intrínseca. Portanto, a dimensão fractal vem sendo utilizada para minimizar significativamente os efeitos da "maldição da dimensionalidade" [Korn 01]. Esse princípio é amplamente explorado em trabalhos de indexação espacial e métrica, e de processamento de consultas. Em [Faloutsos 94] a dimensão fractal do conjunto de dados é utilizada para derivação de fórmulas de estimativa de número de acessos a disco em consultas por abrangência (*Range Queries*) em *R-trees*. A mesma idéia é desenvolvida em [Pagel 00] para consultas aos *k*-vizinhos mais próximos (*k-Nearest Neighbor Queries*). Em [Faloutsos 96] a dimensão fractal é aplicada para estimar o espaço necessário para armazenamento de dados em estruturas *Quadtree*. Em [Böhm 00a] é proposto um modelo de custo para consultas por abrangência e aos *k*-vizinhos mais próximos em espaços de alta dimensionalidade. A dimensão fractal também é aplicada na derivação de fórmulas para estimativa de seletividade em operações de junção espacial [Belussi 98, Faloutsos 00] e consultas por abrangência [Belussi 95]. Em [Böhm 00b] a estrutura proposta para a indexação de conjuntos de dados com alta dimensão de imersão tem seu tamanho de página adaptado dinamicamente de acordo com a distribuição dos dados, avaliada por meio da dimensão fractal do conjunto. Estimativa de custo para consultas por abrangência baseada no conceito de dimensão fractal é proposta em [Traina 02], voltada em particular para a estrutura de indexação métrica *Slim-tree*. Em [Arantes 05], a teoria de fractais fornece embasamento para a definição de um modelo de seletividade para consultas por similaridade complexas em métodos de acesso métricos. Em [Baioco 07] são propostos modelos de custo e de seletividade para dados multimidia.

A dimensão fractal também vem sendo explorada em trabalhos na área de mineração de dados. O algoritmo *Fractal Clustering*, proposto em [Barbará 00], utiliza a dimensão fractal para determinar a similaridade entre objetos e definir agrupamentos. Em [Jiang 02] a análise das propriedades fractais do conjunto de dados é explorada em uma técnica de coleta de amostras que preservem as características de distribuição originais. O trabalho apresentado em [Traina 00] propõe um algoritmo de seleção de atributos baseado na dimensão fractal do conjunto de dados. Uma discussão mais geral da aplicação de conceitos da teoria de fractais em mineração de dados é apresentada em [Traina Jr. 05], assim como técnicas baseadas em fractais aplicadas a processos como detecção de agrupamentos, identificação de *outliers*, redução de dimensionalidade e análise de relacionamentos entre conjuntos de dados. Em [Sousa 07b], a dimensão intrínseca, estimada pela dimensão fractal, é utilizada para identificar correlações entre atributos. Em [Cordeiro 10], a abordagem fractal é utilizada para detecção de agrupamentos em subespaços.

De maneira similar, a dimensão fractal também pode, potencialmente, ser utilizada em algoritmos de análise e processamento de *data streams*, como em estimativas de custo e seletividade de consultas contínuas em *streams* persistentes ou transientes, otimização de consultas, seleção de atributos e previsão de eventos futuros (*forecasting*). Alguns estudos utilizam a dimensão intrínseca, estimada pela dimensão fractal, na área de *data streams* e séries temporais. O trabalho apresentado em [Chakrabarti 02], por exemplo, propõe um método não linear de previsão de elementos futuros em séries temporais, periódicas e caóticas. Em [Kantardzic 04], a dimensão fractal é usada como uma medida de diversidade em portfólios de ações em investimentos financeiros. Em [Sousa 07a] é detalhada uma técnica de cálculo continuado da dimensão fractal em *streams* evolutivas, que pode ser associada a tarefas de de mineração

de *data streams* [Sousa 06b, Romani 09].

As variações da dimensão intrínseca ao longo do tempo indicam variações na correlações entre os atributos que compõem o *data stream*, desse modo, a análise da variação da dimensão fractal ao longo do tempo é uma ferramenta útil na análise de *data streams*. Um modo de viabilizar uma análise continuada sobre conjuntos de dados com as características presentes em *data streams* é por meio da utilização de janelas deslizantes para limitar o escopo de análise sobre a *data stream*, alterando assim o escopo de análise durante o tempo, realizando uma análise continuada do conjunto de dados.

Como mencionado anteriormente, diversos trabalhos utilizam a dimensão fractal para diversas finalidades. Na Seção 3.2 é apresentado o algoritmo *SID-meter*, desenvolvido para realizar o cálculo continuado da dimensão fractal sobre conjuntos de dados com as características presentes em *data streams* multidimensionais. A abordagem proposta no *SID-meter* é utilizada neste trabalho

3.2 *Stream Instrinsic Dimension Meter*

Na análise de *data stream* multidimensional será utilizado uma abordagem semelhante ao Algoritmo *Stream Instrinsic Dimension Meter (SID-meter)* [Sousa 07a]. Esse algoritmo utiliza uma janela móvel baseada em eventos para processar a *data stream*.

O algoritmo SID-*meter* é projetado para estimar a dimensão fractal de *data streams* multidimensionais ao longo do tempo e utiliza uma abordagem baseada no método *Box-Counting*. Desse modo, é necessário definir um hipercubo que englobe os primeiros eventos n_e, para tanto, os primeiros eventos são armazenados temporariamente permitindo assim que o hipercubo inicial seja definido. Até esse momento o algoritmo não é capaz de realizar o cálculo da dimensão fractal de correlações D_2, porém, após esse momento, não é necessário manter nenhum evento em memória e o cálculo da dimensão fractal pode ser realizada a qualquer momento.

Na Figura 3.6 é possível verificar o processo executado pelo pelo *SID-meter*, onde os eventos que compõem o *data stream* chegam continuamente, sendo recebidos pelo método e processados. A partir dos n_e primeiros eventos, as consultas podem ser realizadas a qualquer momento.

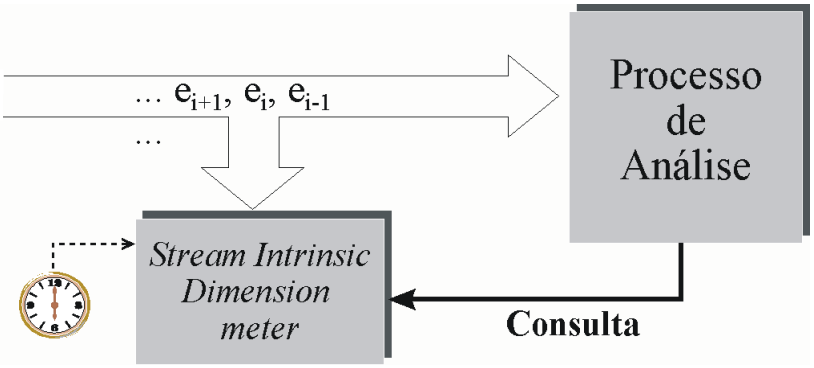

Figura 3.6: Cálculo da Dimensão Fractal em uma *data stream* multidimensional.

Com o hipercubo inicial, uma estrutura de hiper-reticulado é definida para manter a contagem dos pontos em até R reticulados, onde R é o número de pontos utilizados pelo *box-count plot*, tendo seu lado igual a r.

No SID-*meter*, a estrutura utilizada para representar o hiper-reticulado é uma árvore, mantida em memória principal, denominada *box-counting tree*, pois permite realizar a contagem de ocupação ao longo do tempo. A utilização de uma janela deslizante, definida por n_c períodos de contagem onde cada período contém n_e eventos, permite que essa contagem seja atualizada a cada n_e eventos. A cada nível j da árvore, onde $0 \leq j \leq R-1$, as células têm $r_j = r_{j-1}/2$ e para cada célula, são geradas 2^E novas células no nível $j+1$.

Na Figura 3.3 é apresentado um exemplo de *Box-Counting Tree* com com 4 níveis construído sobre um conjunto de eventos. É possível verificar como um hiper-reticulado é construído de modo a englobar todos os eventos do conjunto de dados e também como a árvore de contagem é construída de modo a representar esse hiper-reticulado, mantendo a contagem dos eventos incidentes em cada célula que compõe o hiper-reticulado em seus diferentes níveis.

O *SID-meter* utiliza uma abordagem baseada em janelas deslizantes para definir o escopo de análise da *data stream*. A utilização de janelas deslizantes é importante devido ao fato de que *data streams* são potencialmente infinitas e, portanto, é necessário um mecanismo que permita limitar o escopo de análise da *stream*, além disso, a janela deslizante se desloca a medida que é ocupada pelos novos eventos do conjunto de dados, descartando os eventos mais antigos, permitindo assim manter-se atualizada.

Essa janela é dividida em um número configurável de períodos de contagem n_c, onde cada n_c é formado por um número configurável de eventos n_e. Desse modo, o número de eventos que cada janela mantém é igual a $n_c \times n_e$. Quando a janela inteira está ocupada, a análise é realizada sobre os eventos contidos na janela, os n_e eventos mais antigos são descartados e a estrutura passa a considerar os novos n_e eventos disponíveis na *data stream*.

Na Figura 3.7 é apresentada uma janela deslizante construída sobre um conjunto de dados composto por 3 atributos (a_1, a_2 e a_3). Essa janela deslizante está parametrizada de modo a conter 5 períodos de contagem ($n_c = 5$) sendo que cada período de contagem contém um determinado número de eventos.

É possível verificar que, após preencher completamente a janela deslizante, como apresentado na Figura 3.7 (a), os eventos mais antigos são descartados e os novos eventos recebidos no *data stream* passam a ser considerados, como apresentado na Figura 3.7 (b).

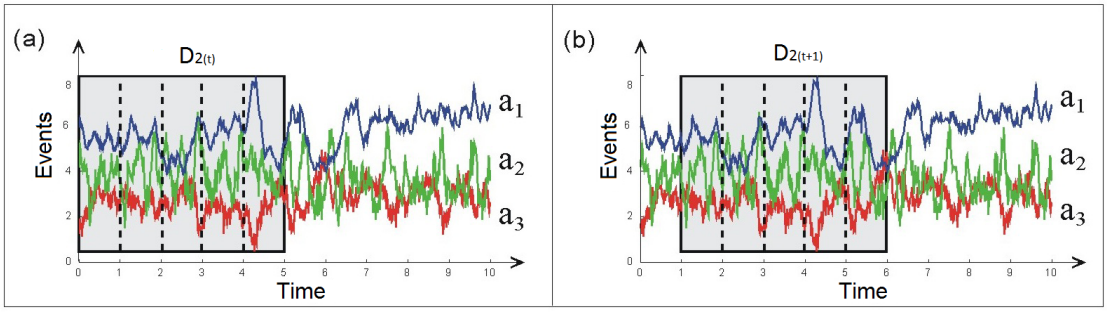

Figura 3.7: Janela deslizante construída sobre 3 atributos contendo 5 períodos de contagem.

Desse modo, em cada nível j da árvore de contagem, está associado um hiper-reticulado com célula de lado igual a r_j, onde os períodos de contagem são armazenados em cada célula por vetores $C[\,]$ de tamanho igual a n_c. Esse vetor é responsável por manter atualizada a contagem para os eventos contidos na janela deslizante definida por n_c e n_e e funciona como uma lista circular por meio da utilização de um contador i_{n_c}. Quando um período expira, o contador i_{n_c} é incrementado para utilizar o próximo período

de contagem e, desse modo, quando $i_{n_c} = n_c$, os eventos mais antigos são descartados e o contador é disponibilizados para realizar a ocupação de novos eventos.

O SID-*meter* realiza o cálculo da dimensão fractal ao final de cada período de contagem n_c, utilizando a Equação 3.3 sobre os eventos contidos na janela deslizante, porém o cálculo da dimensão fractal pode ser realizada a qualquer momento depois que o hipercubo inicial for definido. A contagem de ocupação p_{i,r_j} é realizada por meio do vetor $C[\,]$ a cada nível j sobre cada nó i, como apresentado na Equação 3.4.

$$p_{i,r_j} = \sum_{i_{n_c}=0}^{i_{n_c}<n_c} C_i[i_{n_c}] \tag{3.4}$$

Seguindo essa abordagem, considerando uma janela construída sobre eventos diários, em uma análise em que deve ser realizada sobre as estações do ano, a janela deslizante deve ser dividida de modo que o número de eventos contidos nela sejam referentes as estações do ano, no caso, englobando 3 meses, ou seja, $n_c \times n_e = 90$. Porém, é possível definir diversas configurações de n_c e de n_e que se adequarem a essa configuração, e cada uma delas influência de forma diferente no resultado da análise.

Desse modo, utilizar uma janela com $n_c = 1$ e $n_e = 90$, resultará em uma análise que irá considerar apenas eventos contidos nos últimos 3 meses, porém sem a intersecção entre os meses, sendo a análise realizada a cada 3 meses, por outro lado, janelas configuradas de modo que $n_c = 90$ e $n_e = 1$ também terão uma análise realizada sobre os eventos contidos nos últimos 3 meses, porém com a análise sendo realizada a cada dia. Nesse exemplo, as duas configurações têm granularidade trimestral, porém, a taxa de atualização da primeira janela é trimestral, sendo que a segunda janela tem uma taxa de atualização diária.

A diferença entre essas duas configurações é que a segunda análise irá considerar as intersecções entre as estações, desse modo, em um conjunto de eventos de 180 dias, a primeira configuração de janela irá realizar duas análises, uma para os primeiros 90 eventos e outra para os últimos 90 eventos. Por outro lado, a segunda configuração irá realizar 91 análises, a primeira após preencher completamente a janela (primeiro 90 eventos) e mais uma análise para cada demais evento recebido pelo método.

A granularidade da janela deslizante pode atenuar os resultados sobre os eventos que se deseja analisar, Por exemplo, utilizar janelas temporais com granularidades muito grande, com uma granularidade anual por exemplo, não será muito eficiente se a análise tem como objetivo encontrar alterações na correlação dos atributos que ocorram mensalmente.

Inicialmente, o SID-*meter* realiza o armazenamento dos primeiros n_e eventos em memória principal, permitindo assim definir o hipercubo inicial. Porém, após esse momento, esses eventos são descartados e os demais eventos são utilizados apenas pare realizar a atualização das contagens de ocupação. Porém, os demais eventos após os n_e primeiros eventos podem extrapolar o espaço de representação do hipercubo inicial.

Desse modo, o algoritmo tem mecanismos para comportar eventuais expansões e contrações do espaço representado pela árvore de contagem. Portanto, são quatro movimentos possíveis realizados no espaço de representação.

- **Ocupação de regiões:** Ocorre quando um novo evento incide em uma célula anteriormente vazia da estrutura da árvore de contagem. Desse modo um novo nó é alocado no nível correspondente da árvore.

- **Liberação de regiões:** Ocorre quando uma célula anteriormente ocupada fica vazia devido a eventos antigos descartados.

- **Expansão do espaço de representação:** Ocorre quando um novo evento incide fora do espaço representado pela árvore de contagem. O movimento de expansão requer uma atualização das dimensões do hipercubo, de modo a englobar o novo evento que não está dentro dos limites. Esse processo é realizado criando um nível novo acima da árvore de contagem, cujo lado será $r \times 2$. Caso o novo hipercubo ainda não englobe o novo evento, esse processo é repetido até que o hipercubo englobe o novo evento.

- **Contração do espaço de representação:** Esse movimento ocorre quando apenas uma célula filha no segundo nível da árvore de contagem é ocupada, desse modo, essa célula passa a ser a raiz da árvore de contagem. Na prática, isso representa uma redução do representado pela árvore de contagem.

Esses movimentos, apresentados na Figura 3.8, permitem que a estrutura utilizada pelo SID-*meter* se adeque a mudanças na distribuição do conjunto de dados analisados. Desse modo, esse algoritmo é uma solução com um custo computacional viável para a análise temporal da variação da correlação de *data streams* multidimensionais, considerando que as características presentes nesses conjuntos de dados fazem com que métodos de análise se adequem a dados que podem ser potencialmente infinitos, além de que a distribuição dos eventos que compõem esses dados podem mudar ao longo do tempo.

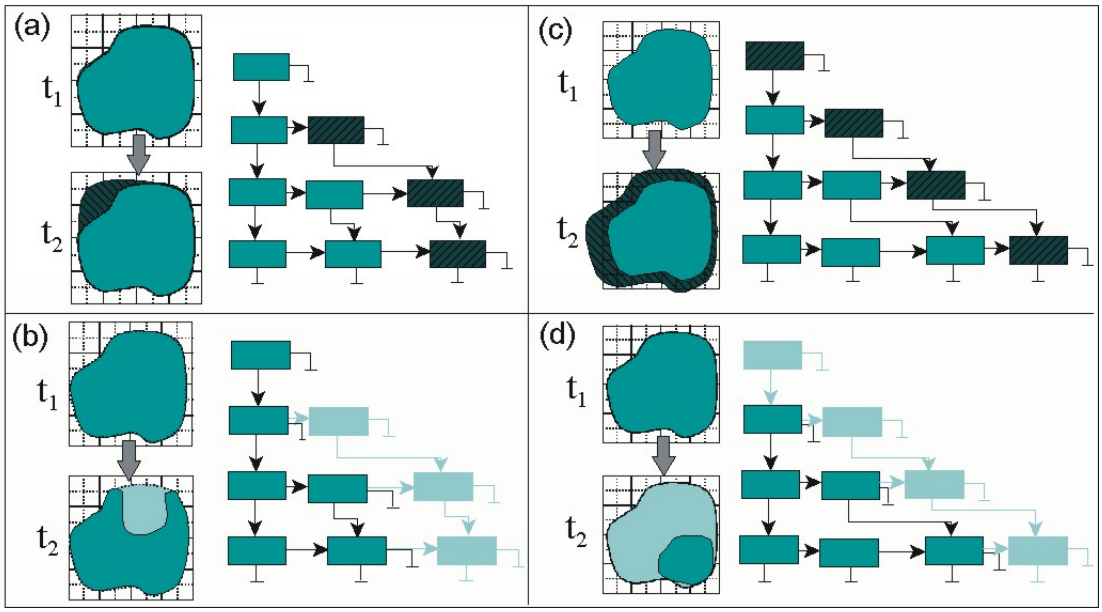

Figura 3.8: Movimentos realizados sobre a estrutura *Box-Counting Tree*: (a) Ocupação de regiões; (b) Liberação de regiões; (c) Expansão do espaço de representação; (d) Contração do espaço de representação [Sousa 06a].

Considerações Finais

Neste Capítulo foram apresentadas técnicas essenciais para o cálculo da dimensão fractal sobre *data streams* multidimensionais, essas técnicas permitem realizar uma análise temporal da variação da correlação entre os atributos que compõem a *stream* ao longo do tempo.

O Capítulo 4 apresenta o método STB-*meter* que tem como objetivo realizar o cálculo da dimensão fractal de correlação sobre *data streams* multidimensionais, considerando também aspectos espacial dos dados analisado. Para tanto, foi utilizada uma abordagem semelhante a do método *SID-meter*, ou seja, que utiliza uma *box-counting tree* para realizar o cálculo continuado da dimensão fractal de correlação D_2, porém, utilizando características presentes em estruturas de dados hierárquicas, considerando também aspectos espaciais do conjunto de dados. Além disso, a abordagem de janelas deslizantes foi estendida, de modo a permitir que diversas granularidades temporais sejam consideradas simultaneamente.

Capítulo 4

Spatio-Temporal Behaviour Meter

Tarefas de análise e mineração em *data streams* tratam fluxos de dados contínuos e potencialmente infinitos. Em geral, o interesse está em extrair padrões e conhecimento a partir dos dados recentes, atualizando esse conhecimento ao longo do tempo e analisando a evolução dos dados. Logo, o aspecto temporal está inerentemente associado o *data streams*.

Esses conjuntos de dados também podem possuir atributos referentes a informações espaciais que são gerados continuamente ao longo do tempo. Desse modo, é possível utilizar os aspectos espaciais do conjunto de dados, assim como os atributos não espaciais, na tarefa de descoberta de conhecimento e possivelmente auxiliar o especialista do domínio em sua tarefa de tomada de decisão.

Diversos conjuntos de dados apresentam essa característica, como sensores distribuídos em uma área visando obter a temperatura e precipitação ao longo do tempo, gerando assim uma série de medições compostas de variáveis espaciais (localização geoespacial do sensor) e não espaciais (medidas de temperatura e precipitação). Nesses conjuntos de dados, uma análise espaço-temporal pode auxiliar o especialista do domínio da aplicação a compreender o comportamento dos atributos climáticos considerando diversas localidades diferentes em várias escalas temporais. Essa análise permite analisar, por exemplo, como um evento extremo como o *El Niño* afeta o clima de diferentes sub-regiões, como o Sul e Sudeste do Brasil.

Nesse contexto, o método *Spatio-Temporal Behaviour Meter (STB-meter)* foi desenvolvido de modo a viabilizar a análise de *data streams* multi-dimensionais com atributos espaciais e não espaciais, com o principal objetivo de permitir o cálculo contínuo da dimensão fractal de correlação sobre os atributos que compõem o *stream* considerando múltiplas sub-regiões do espaço assim como múltiplas granularidades temporais, permitindo assim, analisar a correlação entre os atributos não espaciais em diferentes granularidades espaciais.

A análise desses resultados auxilia o especialista do domínio a compreender diversos aspectos relacionados a correlação entre os atributos não espaciais do conjunto de dados, possibilitando verificar o impacto causado por eventos extremos nas correlações entre temperatura e precipitação em diferentes regiões do pais.

4.1 Técnicas fundamentais utilizadas pelo método *STB-meter*

Nessa seção são apresentadas as soluções adotadas no método *STN-meter* para realizar a representação dos atributos que compõem o *data stream* multi-dimensionais de modo a permitir que uma análise conjunta dos aspectos espaciais e não espaciais seja realizada.

O método *STB-meter* cria dois espaços de representação de modo que os atributos espaciais e os atributos não espaciais sejam englobados em espaços distintos. Essa distinção é realizada pois cada subconjunto de atributos (espaciais e não espaciais) é utilizado com finalidades distintas para o método. O conjunto dos atributos espaciais é utilizado para definir as sub-regiões do espaço que serão consideradas para realizar o cálculo da dimensão fractal D_2 e, para tanto, utilizam-se estruturas hierárquicas de dados para realizar tal representação. Por outro lado, o conjunto de atributos não espaciais é utilizado para realizar a contagem de pontos incidentes em cada sub-região definida pela estrutura de dados hierárquica.

A representação dos atributos não espaciais é realizada utilizando uma abordagem semelhante ao *SID-meter*, descrito na Seção 3.2, ou seja, baseada em uma *Box-counting tree*. Essa estrutura permite que a correlação dos atributos não espaciais seja calculada realizando apenas uma leitura de cada evento que compõe o conjunto de dados.

Além disso, o *SID-meter* utiliza janelas deslizantes para limitar o escopo de análise da *stream*. Essa abordagem foi estendida de forma que a janela deslizante considere um intervalo de valores para os parâmetros n_c (número de períodos de contagem) e n_e (número de eventos por período de contagem). Por exemplo, considerando uma *data stream* composta por eventos diários, é possível configurar n_c no intervalo de 1 a 4 períodos de contagem e n_e com intervalo de 30 a 90 dias. Como resultado, teremos uma única janela armazenando eventos que permitem realizar análises com a granularidades temporais contidas em {30, 60, 90, 120, 150, 180, 210, 240, 270, 300, 330 e 360} dias.

Essa extensão se dá por meio da utilização de vetores para armazenar a contagem de eventos incidentes na janela temporal de cada célula da *Box-counting tree*. No exemplo anterior, considerando um conjunto com 2 atributos não espaciais, cada célula da *box-counting tree* será subdividida em 4 subregiões.

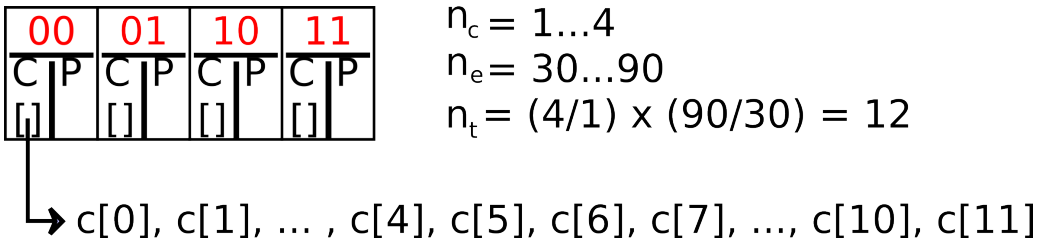

Figura 4.1: Representação da Janela Deslizante estendida nos nós que compõe a *Box-counting tree*.

Na Figura 4.1 é apresentado um nó e sua respectiva janela deslizante configurada com n_c no intervalo de 1 a 4 e n_e com intervalo de 30 a 90. Na Figura 4.2 é apresentada a mesma janela sobre uma *data stream* bidimensional, sendo que o contador $C[\,]$ realiza a contagem de eventos incidentes.

A ideia principal dessa extensão é permitir que diversas granularidades temporais sejam consideradas de forma simultânea, visando aumentar a capacidade de análise do especialista do domínio ao aumentar as possibilidades de análise sobre o mesmo conjunto de dados.

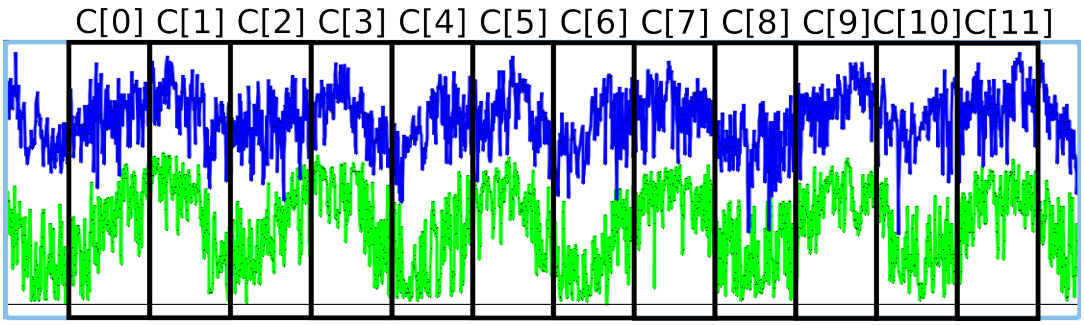

Figura 4.2: Representação da Janela Deslizante estendida sobre um *data stream* bidimensional.

Na implementação, para realizar a análise sobre a janela deslizante, passa-se a considerar um ponteiro referente ao período de contagem atual i_{nc} que irá receber os novos eventos. No exemplo anterior, as granularidades possíveis vão de 30 a 360 dias e, desse modo, uma análise que utiliza uma granularidade de 180 dias deve considerar os 6 últimos períodos de contagem a partir do ponteiro i_{nc}. Essa abordagem é ilustrada na Figura 4.4 onde pode-se verificar como uma análise realizada sobre a janela deslizante estendida considera apenas o conjunto de dados referente aos 6 períodos de contagem mais recentes, ou seja, aos últimos 180 dias. Para tanto, a partir do contador i_{nc}, utiliza-se as ultimas 6 posições do vetor de contagem $C[\,]$, como ilustrado na Figura 4.3.

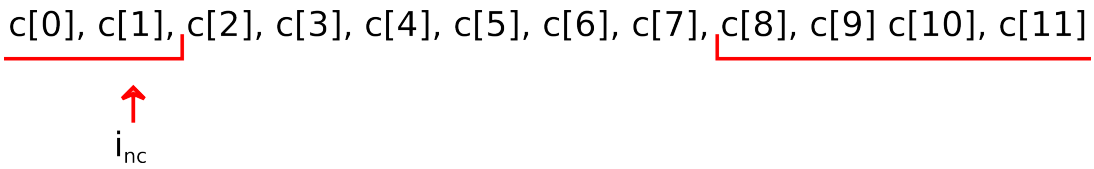

Figura 4.3: Períodos de contagem utilizados na contagem de eventos considerando os últimos 180 dias.

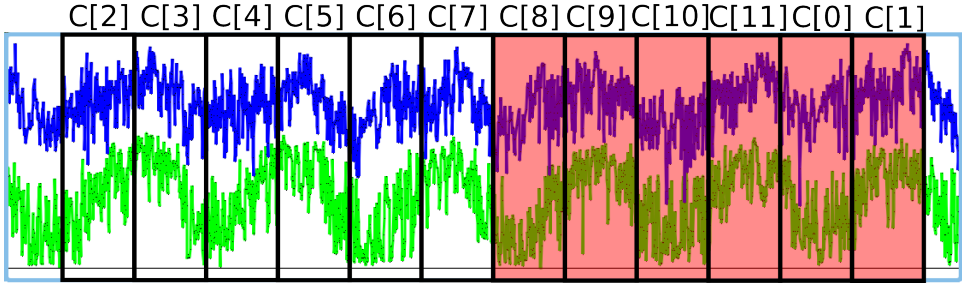

Figura 4.4: Períodos de contagem utilizados na contagem de eventos considerando os últimos 180 dias.

Outra característica relevante é a representação da informação espacial. O *STB-meter* utiliza uma estrutura de dados hierárquica para realizar essa representação dos atributos espaciais do conjunto de dados. Estruturas de dados hierárquicas tem como uma de suas principais características a subdivisão recursiva do espaço representado [Samet 84]. Essa característica é fundamental para o método *STB-meter* para viabilizar uma análise que leve em consideração aspectos espaciais e temporais do conjunto de dados de forma simultânea. Assim, no método *STB-meter* foi criada uma estrutura de dados chamada *STB-tree*, que integra aspectos presentes na *box-counting tree* estendida com aspectos presentes em estruturas de dados hierárquicas, além de considerar a extensão realizada a abordagem de janelas deslizantes de

modo que os parâmetros que as definem seja um intervalo de valores, permitindo a análise de múltiplas granularidades simultaneamente, tanto espacial quanto temporal..

Desse modo, dado um *data stream* composto pelos atributos espaciais $a_s = (a_{S_1}, ..., a_{N_S})$ e também pelos atributos não espaciais $a_s = (a_{V_1}, ..., a_{N_V})$ onde $N_V, N_S > 0$. Constrói-se uma representação utilizando uma estrutura de dados hierárquica S que realiza uma subdivisão do espaço de baseada nos atributos a_s de acordo com sua regra de construção.

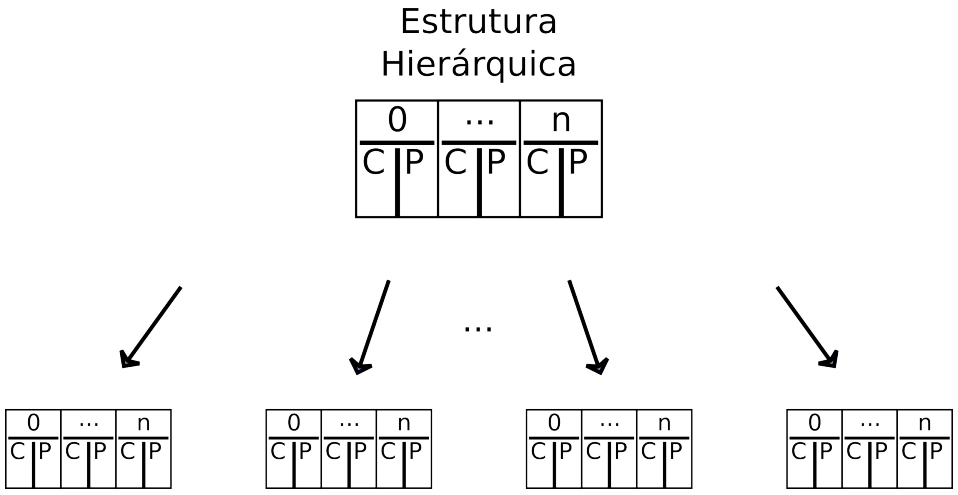

Figura 4.5: Estrutura de dados hierárquica.

Na Figura 4.5 é apresentada uma abstração de uma estrutura hierárquica que realiza a divisão do espaço de representação em n sub-regiões, de acordo com a regra de divisão recursiva adotada pela mesma. Pode-se observar que a cada sub-região estão associados um ponteiro P referente ao próximo nível na estrutura hierárquica e um vetor $C[\,]$ responsável por manter a contagem de pontos incidentes naquela sub-região. Além disso, cada nó contém um índice que o identifica.

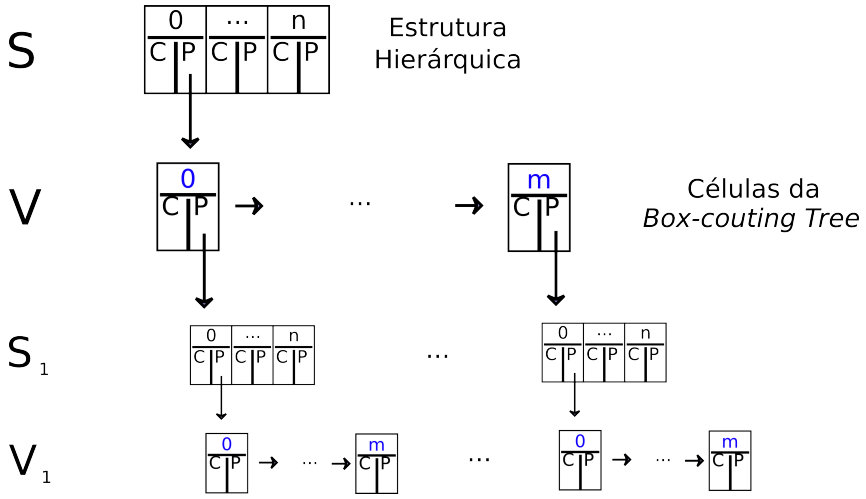

Figura 4.6: Estrutura da *STB-tree* : Integração de estrutura de dados hierárquica com a árvore de contagem estendida.

De maneira semelhante, sobre os atributos não espaciais a_v é definido um espaço de representação V por meio da utilização de um hiperreticulado com N_V dimensões, seguindo a abordagem do *Box-counting*

tree (árvore de contagem). A integração da estrutura de dados hierárquica com a *Box-counting tree* estendida define a *STB-tree*, ilustrada na Figura 4.6.

Pode-se observar que o primeiro nível da *STB-tree* é composto pelo primeiro nível da estrutura de dados hierárquica S, representando todo o espaço definido pelos atributos espaciais do conjunto de dados. No segundo nível da *STB-tree* está presente a estrutura utilizada para realizar a contagem dos pontos definidos pelos atributos não espaciais, ou seja, o primeiro nível da *Box-counting tree* V.

Para cada célula de V, é atribuído o próximo nível da estrutura da dados hierárquica, ou seja, o nível S_1 seguindo a regra de subdivisão do espaço adotada pela estrutura. Desse modo, o terceiro nível da *STB-tree* é formado pelos nós da estrutura de dados hierárquica associado a cada célula da *Box counting tree*.

Sobre cada um dos nós $S_{1_1}, ..., S_{1_I}$, tem-se um conjunto de eventos do *data stream* incidentes em cada sub-região. Desse modo, para cada sub-região de S há um hiperreticulado para realizar a contagem dos atributos não espaciais do *stream*. O quarto nível da *STB-tree* é formado pelos nós do segundo nível da *box-counting tree* construída sobre um subconjunto dos eventos da *stream* definidos pelas células $V_{1_1}, ..., V_{1_J}$, e assim sucessivamente.

Esse estrutura cresce até que a profundidade k seja atingida, onde cada nível é composto por um sub-nível que representa a informação espacial do conjunto de dados e outro que realiza a contagem das informações não espaciais.

O principal objetivo do *STB-tree* é permitir que seja possível obter a dimensão fractal de correlação ao longo do tempo, utilizando a abordagem do *Box-counting* sobre cada sub-região de S, considerando diversas granularidades temporais e maximizando assim a capacidade de análise do especialista do domínio.

4.2 Implementação Inicial

A estrutura *STB-tree* foi concebida de modo que diversas estruturas de dados hierárquicas possam ser utilizadas em sua implementação, tais como a *K-D-B-tree* [Robinson 81], *R-tree* [Guttman 84] e *Quad-tree* [Samet 84]. Essa abordagem foi adotada pelo fato de que a regra de divisão recursiva adotada pela estrutura de dados hierárquica influencia no modo que as sub-regiões do espaço são criadas a partir dos atributos espaciais do conjunto de dados.

Nesse cenário, é possível realizar diversas implementações da *STB-tree*, considerando diversas estruturas de dados hierárquicas. Neste trabalho adotada a *Regions Quadtrees* como estrutura de dados hierárquica para realizar a representação dos atributos espaciais. Sobre essa estrutura, duas operações são de fundamental relevância no escopo do trabalho de modo a viabilizar sua utilização da *STB-tree*:

1. Construção da estrutura de dados: A *region quadtree* decompõe o espaço representado de E dimensões na ordem de 2^E subconjuntos de mesmo tamanho para cada nível, até que cada célula contenha um limite máximo pré-definido de elementos [Samet 84]. Desse modo, será utilizada uma estrutura de árvore para representar essa estrutura de dados. Essa decomposição será realizada até que se chegue aos nós folha, onde a altura da árvore que representa essa estrutura será definida previamente pelo usuário.

2. Buscas: Após a construção da estrutura de dados, a operação de busca de subconjuntos será constante. Nessa estrutura, a busca é realizada de maneira similar à busca realizada sobre árvores

binárias [Gaede 98], ou seja, a cada nível, um dos nós-filho será eleito para realizar a nova busca. A seleção de regiões é realizada de maneira similar, a diferença é que a busca termina quando se atinge um determinado nível da árvore, selecionando assim um subconjunto de pontos.

Considerando então uma *STB-tree* que utiliza a *Regions Quadtree* como estrutura de dados hierárquica e um conjunto de evento $e_1, e_2, e_3, ...$ onde cada evento e_i é composto por um conjunto de atributos $A = (a_1, a_2, ..., a_{N_S}, a_{N_S+1}, ..., a_{N_S+N_V})$ onde os atributos $a_1, ..., a_{N_S}$ são referentes a informações espaciais e os atributos $a_{N_S+1}, ..., a_{N_S+N_V}$ são referentes a informações não espaciais, a construção da *STB-tree* está descrita no Algoritmo 4.1 e ocorre do seguinte modo:

1. Inicialmente, os primeiros n_e eventos são armazenado temporariamente. O objetivo desse processo é utilizar esse subconjunto de dados para definir o espaço de representação inicial, tanto para os atributos espaciais como para os atributos não espaciais. Esse processo pode ser verificado nas linhas 2 e 16, onde o subconjunto temporário é definido e também onde os eventos do *data stream* são armazenados temporariamente.

2. O segundo passo é verificar se o conjunto temporário com n_e eventos já foi preenchido e, a partir desse momento, o *STB-tree* é configurado de modo que seus espaços de representação englobem tantos os atributos espaciais como os não espaciais. Esse processo é apresentado nas linhas 8 a 14 do algoritmo 4.1.

   Inicialmente define-se os espaços de representação da *quadtree* e da árvore de contagem, como apresentado nas linhas 9 e 10 respectivamente. Em seguida, cada evento armazenado temporariamente é contabilizado pela *STB-tree*, como realizado nas linhas 11 a 13.

3. Finalmente, após a contagem dos primeiros n_e eventos, a estrutura passa a receber os demais eventos de forma contínua, como apresentado na linha 7 , sendo possível realizar contagens na estrutura a partir desse ponto.

O Algoritmo 4.1 realiza a parametrização inicial da estrutura *STB-tree*. Nesse processo, os n_e primeiros eventos recebidos são armazenados temporariamente de modo a obter as informações necessárias para configurar o espaço de representação, tanto da estrutura hierárquica de dados utilizada, como também da *Box-counting tree* estendida, utilizada para realizar a contagem dos eventos incidentes.

A configuração do espaço de representação utilizado na *quadtree* consiste em detectar a maior distância r_s dentre os atributos espaciais e utilizar essa distância como tamanho do espaço representado pela estrutura de dados hierárquica. A abordagem utilizada na configuração do espaço de representação utilizada pela *box couting tree* realiza um processo semelhante, porém considerando a maior distância r_s dentre os atributos não espaciais do conjunto de dados.

Após a configuração dos espaços de representação, o algoritmo passa a receber e processar os demais eventos recebidos, sendo possível realizar as consultas necessárias para o cálculo da dimensão fractal de correlação para diferentes sub-regiões do espaço.

O processamento de cada evento e na *STB-tree* ocorre do seguinte modo: para cada evento e, separam-se os atributos em dois subconjuntos, onde a_s é constituído pelos atributos espaciais e a_v é constituído pelos atributos não espaciais, como apresentado nas linhas 1 e 2 do Algoritmo 4.2. Mantém-se

Algoritmo 4.1 Construção da *STB-tree*

Entrada: A: conjunto de eventos.
k: profundidade desejada.
n_c: número de períodos de contagem.
r_c: intervalo de n_c.
n_e: número de eventos por períodos de contagem.
r_e: intervalo de n_e.
Saída: S: *STB-tree*.

1: Número de eventos recebido $n \leftarrow 0$
2: Cria conjunto t de atributos com tamanho n_e
3: Cria *STB-tree* s contendo uma *Regions Quadtree* q e um *Box Counting Tree* estendida b utilizando janelas deslizantes com $(r_e \times (n_c + r_c))$ períodos de contagem e n_e eventos por período de contagem e com k níveis,
4: **enquanto** A **not** $\emptyset$ **faça**
5: $a \leftarrow$ Recebe próximo evento de A
6: **se** $n > n_e$ **então**
7: RealizaContagem(s, a)
8: **senão se** $n = n_e$ **então**
9: Configura o espaço de representação de q utilizando a informação espacial dos atributos contidos em t
10: Configura o espaço de representação de b utilizando a informação dos atributos não espaciais contidos em t
11: **para** cada evento x contido em t **faça**
12: RealizaContagem(s, x)
13: **fim para**
14: RealizaContagem(s, a)
15: **senão**
16: Adiciona em t o evento a
17: **fim se**
18: $n \leftarrow n + 1$
19: **fim enquanto**

uma referência do primeiro nível da estrutura de dados hierárquica da *STB-tree* e essa referência é atualizada à medida que os níveis inferiores da estrutura são acessados. Esse acesso se dá de acordo com as sub-regiões das representações espacial (*Quadtree*) e não espacial (*Box Counting Tree*) em que a_s e a_v incidem, respectivamente, como apresentado no bloco 5 a 11 do Algoritmo 4.2.

A cada novo nível acessado, realiza-se o incremento da contagem mantida por cada célula onde o evento incide. Contudo, esse processo deve levar em consideração a extensão realizada sobre a janela deslizante. Para tanto, a estrutura *STB-meter* mantém a contagem do número de eventos já recebidos, permitindo assim gerenciar qual período de contagem está sendo utilizado no momento de realizar o incremento.

No Algoritmo 4.3 é apresentado esse processo. Pode-se verificar que inicialmente, o número total de períodos de contagem parametrizados na *STB-tree* (como descrito no Algoritmo 4.1) e o período de contagem atual são utilizado para:

1. Realizar o incremento da contagem da célula (linhas 3 e 4 do algoritmo 4.3) considerando o período de contagem.

Algoritmo 4.2 Procedimento *RealizaContagem*

Entrada: *S*: *STB-tree* com os espaços de representação já definidos.
e: evento a ser inserido na estrutura.
Saída: *S*: *STB-tree*.

1: $a_s \leftarrow$ Recebe atributos espaciais de e
2: $a_v \leftarrow$ Recebe atributos não espaciais de e
3: $leaf \leftarrow$ Estrutura Hierárquica (*quadtree*) utilizada em S
4: $i \leftarrow 0$
5: **para** $i \leq$ número de níveis de S **faça**
6: $childLeaf \leftarrow$ Busca sub-região de $leaf$ onde a_s incide.
7: $childCell \leftarrow$ Busca sub-região de $childLeaf$ onde a_v incide.
8: Incrementa a contagem de $childCell$.
9: $leaf \leftarrow$ sub-regiões associadas a $childCell$.
10: $i \leftarrow i+1$
11: **fim para**

Algoritmo 4.3 Procedimento *IncrementaContagem*

Entrada: *C*: Célula da árvore de contagem.
S: *STB-tree*

1: $n_etotal \leftarrow$ Total de eventos já recebidos em S
2: $n_catual \leftarrow$ período de contagem atual S
3: $C[n_catual] \leftarrow C[n_catual]+1$
4: $n_etotal \leftarrow n_etotal+1$
5: Numero total de eventos já recebidos de $S \leftarrow n_etotal$
6: **se** $n_etotal \bmod n_e = 0$ **então**
7: $n_catualS \leftarrow n_catual+1$
8: período de contagem de $S \leftarrow n_catual+1$
9: $C[n_catual] \leftarrow 0$
10: **fim se**

2. Atualizar o período de contagem atual mantido pela estrutura *STB-tree*, ou seja, caso o evento recebido preencha completamente o período de contagem atual, deve-se descartar os n_e eventos mais antigos contidos na janela e passar a considerar os novos eventos recebidos. Esse processo está descrito no bloco formado pelas linhas 6 a 10 do algoritmo 4.3, onde a referência do período de contagem atual da estrutura *STB-tree* é atualizado, atribuindo zero ao período de contagem mais antigo e atualizando *STB-tree*.

O processo descrito pelos Algoritmos 4.1, 4.2 e 4.3 permite que seja realizada a construção da *STB-tree* assim como o processamento de eventos de um *data stream* com informações espaciais e não espaciais. Além disso, a estrutura mantém a contagem de eventos incidentes em diferentes granularidades espaciais, por meio da utilização da estrutura de dados hierárquica, e também diferentes granularidades temporais, realizada através da extensão realizada sobre as janelas deslizantes.

Por exemplo, considere um *data stream* formado pelos eventos apresentados na Tabela 4.1, onde os dois primeiros atributos são referentes a informações espaciais do conjunto de dados, e também uma *STB-tree* que utiliza a *region quadtree* como estrutura hierárquica e configurada para ter 3 níveis de profundidade.

Tabela 4.1: Exemplo de sequência de eventos com compõem um *data stream* multidimensional.

Eventos	Atributos
e_1	{0,0,4,8}
e_2	{4,4,5,8}
e_3	{7,7,8,0}
e_4	{8,6,6,3}
e_5	{7,5,4,5}

Para ilustrar o processo de construção definido no Algoritmo 4.1, todos os eventos presentes na Tabela 4.1 serão considerados na etapa de construção da *STB-tree*, preenchendo assim a janela $C[0]$. Inicialmente, deve-se verificar o espaço de representação necessário para englobar os eventos contidos na janela inicial, tanto para a informação espacial como para a informação não espacial. Nesse caso, considerando que os dois primeiros atributos são referentes a informações espaciais, teremos o espaço de representação inicial definido do modo apresentado na Figura 4.7.

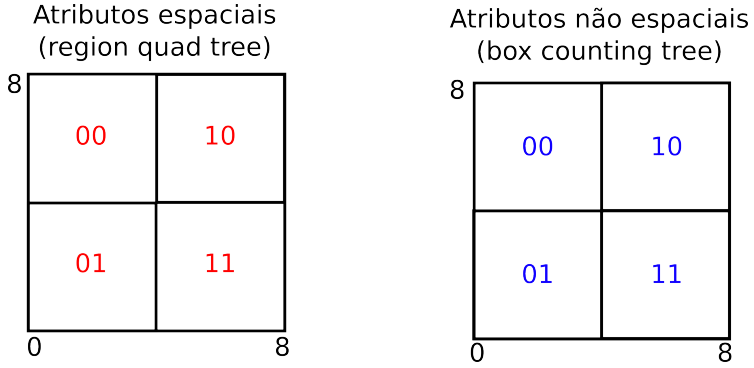

Figura 4.7: Espaço de representação para os atributos espaciais (esquerda) e não espaciais (direita).

Após definidos os espaços de representação, a estrutura começa a receber os eventos que compõem o *data stream*. Nessa etapa, a cada novo evento processado, a *STB-tree* utiliza seus atributos para selecionar qual sub-região da representação espacial e não espacial o evento incide.

Desse modo, após o processamento do evento e_1, a estrutura *STB-tree* passa a ter a forma apresentada na Figura 4.8. É possível observar que o evento e_1 ocupa diferentes sub-regiões da representação espacial e não espacial. Isso ocorre devido ao fato de que as estruturas internas que compõem o *STB-tree* levam em consideração aspectos distintos de cada evento. A estrutura de dados hierárquica leva em consideração apenas os atributos $a_s = \{0,0\}$ do evento e_1 e utiliza um nó parametrizado para representar um espaço variando de 0 a 8 em ambas as dimensões, de modo que a sub-região em que $a_s = \{0,0\}$ incide é representado pelo nó com índice igual a 01 no primeiro nível da estrutura de dados hierárquica.

Por outro lado, a *Box-counting tree* leva em consideração os atributos $a_v = \{4,8\}$ para selecionar a sub-região onde o evento incide, sendo a sub-região em que os atributos não espaciais incidem representada pelo índice 00 no primeiro nível da *Box-counting tree*. Esse processo é executado iterativamente até que a profundidade k desejada na *STB-tree* seja alcançada. Além disso, é possível notar na Figura 4.8 que cada representação mantém uma contagem de eventos incidentes em cada sub-região além de

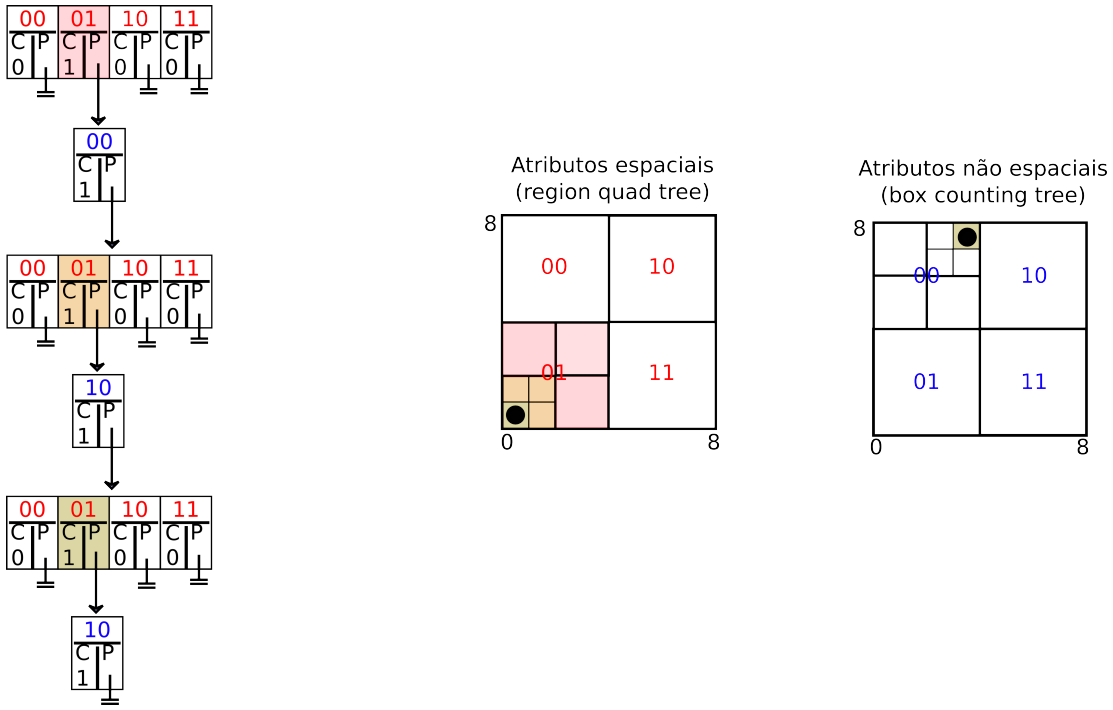

Figura 4.8: *STB-tree* após o processamento do evento $e_1 = \{0,0,4,8\}$.

indexar a região de acordo com sua posição na estrutura.

No processamento do evento e_2, os atributos espaciais $a_s = \{4,4\}$ incidem na mesma sub-região que o evento e_1 no primeiro nível da estrutura de dados hierárquica da *STB-tree*, indexado com o valor 01. Logo, a contagem desse nó da estrutura hierárquica será igual a 2. Esse nó contém uma célula da árvore de contagem associada, referente à sub-região 00 da *box counting tree*, porém, como os atributos não espaciais $a_v = \{5,8\}$ de e_2 incidem em uma sub-região diferente de e_1, é necessário criar uma nova célula para representar a sub-região em que os atributos não espaciais de e_2 incidem. Nesse ponto, o nó da estrutura hierárquica passa a referenciar duas células da *box couting tree*. Na Figura 4.9 é possível verificar como a estrutura da *STB-meter* se comporta após o processamento do evento e_2.

Seguindo o processo apresentado nos Algoritmos 4.1, 4.2 e 4.3 para os eventos e_3, e_4 e e_5 da Tabela 4.1, a *STB-tree* terá a estrutura apresentada na Figura 4.10.

Após a construção inicial da estrutura, ela passa a receber e processar os eventos do *data stream* de forma contínua, sendo possível realizar consultas na estrutura para verificar a contagem de eventos incidentes para diferentes valores de r representados pela *box counting tree*. Para tanto, é necessário realizar o somatório de eventos da janela deslizante incidentes em cada célula de modo que seja possível utilizar a Equação 3.3 para realizar o cálculo da Dimensão Fractal de Correlação D_2.

Desse modo, dada uma janela deslizante configurada com número de períodos de contagem variando um intervalo de n_c a $n_c \times r_c$, sendo r_c o multiplicador de n_c que define o intervalo de valores de n_c considerados pela janela, e cada período de contagem contendo quantidade de eventos no intervalo de n_e a $n_e \times r_e$, onde r_e é o multiplicador de n_e que define o intervalo de valores de n_e considerados pela janela. A janela deslizante será configurada com $n_t = n_c \times r_c \times r_e$ períodos de contagem, isto é, o vetor alocado para representa a janela deslizante será $C[n_t]$, onde cada período irá conter n_e eventos. Em um exemplo prático, dados os parâmetros de números de períodos de contagem $n_c = 1$ com $r_c = 2$ e os parâmetros

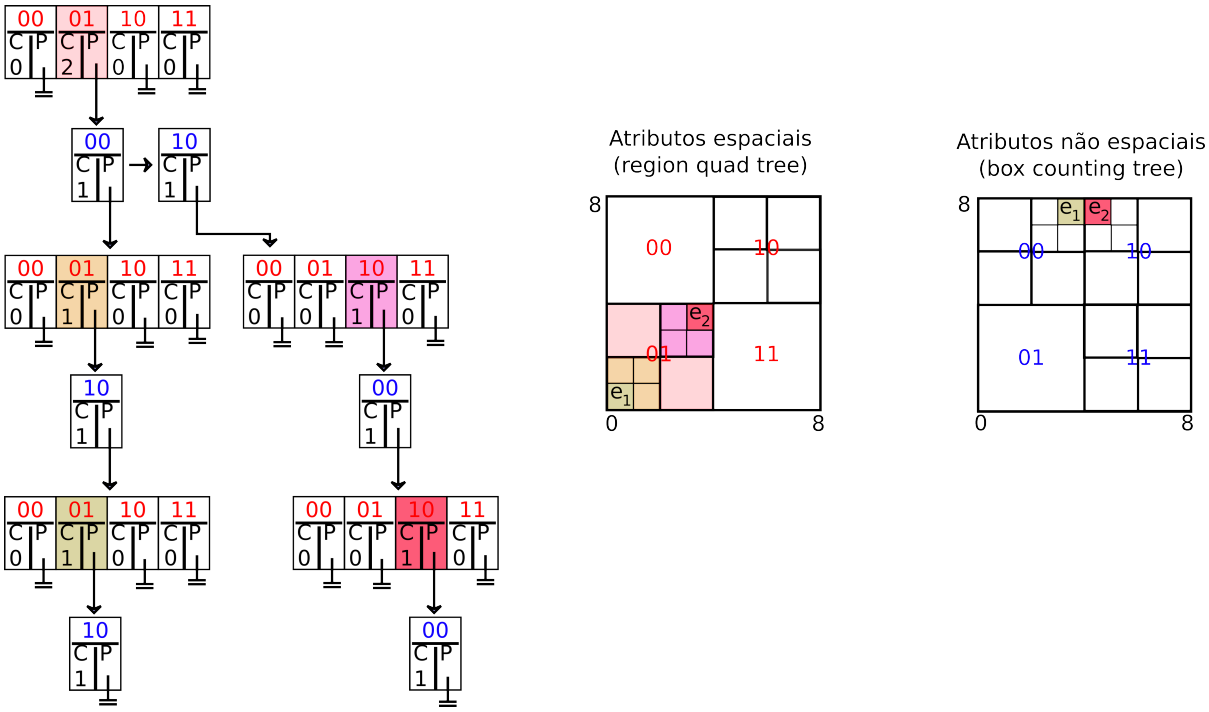

Figura 4.9: *STB-tree* após o processamento do evento $e_2 = \{4,4,5,8\}$.

de números de eventos por período de contagem $n_e = 30$ com intervalo $r_e = 3$, a janela deslizante será configurada com $1 \times 2 \times 3$ períodos de contagem onde cada período é referente a 30 eventos. Essa configuração fará com que seja possível realizar consultas com as granularidades $30, 60, 90, 120, 150$ e 180.

Nesse cenário, o processo apresentado no Algoritmo 4.4 permite obter a contagem de eventos incidentes em cada célula considerando a extensão realizada na janela deslizante. Portanto, a partir do contador atual i_{nc} mantido pela *STB-tree* e dado uma determinada granularidade g definida pelo usuário, o Algoritmo realiza a contagem a partir de i_{nc} considerando g períodos anteriores.

O processo descrito pelos algoritmos 4.1, 4.2 e 4.3 permite que a *STB-tree* mantenha a contagem de pontos incidentes, já considerando os aspectos espaciais e não espaciais dos eventos que compõem o conjunto de dados.

Na Figura 4.11 é possível observar como o Algoritmo 4.4 considera diferentes períodos de contagem para realizar o cálculo de D_2 para diferentes granularidades, partindo do contador $C[i_{nc}]$ mantido pela estrutura *STB-tree*.

Nesse cenário, a utilização de memória pela estrutura *STB-meter* é proporcional ao número total de períodos de contagem n_t considerando o número de nós criados por nível. Esse número é diretamente relacionado a estrutura de dados hierárquica utilizada. Desse modo, a implementação que utiliza a *Region Quadtree* irá dividir o espaço de representação em 2^{N_S} sub-regiões, e do mesmo modo, a representação da *box-counting tree* irá dividir o espaço representado em 2^{N_V}. Portanto, dado um número de níveis k, no pior caso onde cada evento incide em uma região separada da representação espacial e não espacial em todos os níveis de refinamento, a utilização de memória é da ordem de $O(n_t \times (2^{N_S} \times 2^{N_V})^k)$. Portanto, para realizar o cálculo da dimensão fractal D_2 considerando todo o espaço de representação da *STB-tree*,

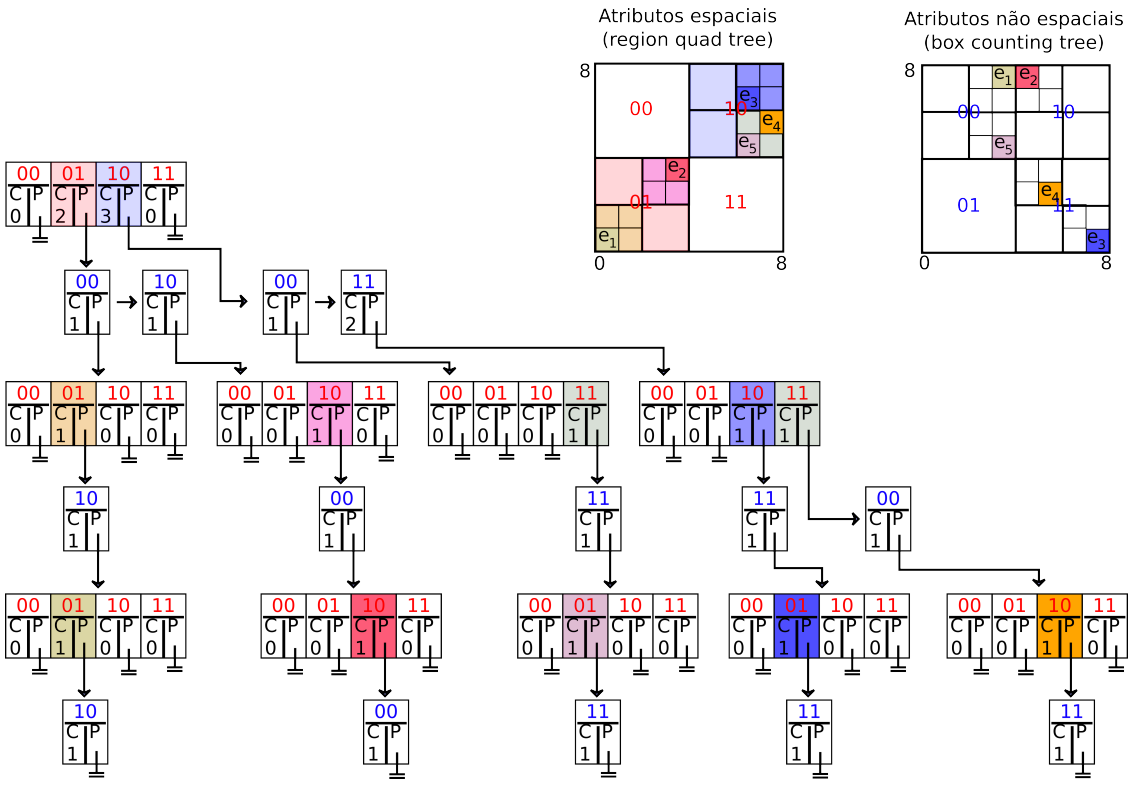

Figura 4.10: *STB-tree* após o processamento de todos os eventos da Tabela 4.1.

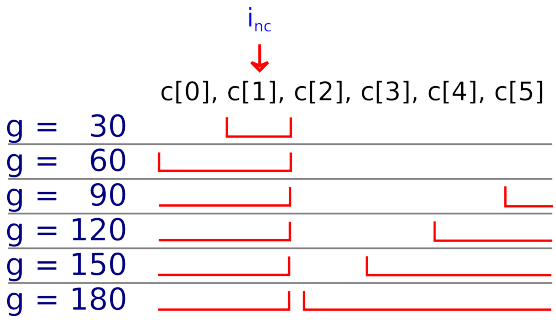

Figura 4.11: Exemplo de granularidades possíveis para o cálculo de eventos incidentes em uma célula da *STB-tree*.

todos os nós da estrutura deverão ser acessados assim como todos os períodos de contagem, de modo que a complexidade do cálculo da dimensão fractal no pior caso é idêntica à ordem de uso de memória.

O processo apresentado para construção da *STB-tree* permite que os eventos sejam recebidos e processados de forma contínua. Essa característica é importante devido à natureza de *Data Streams* evolutivas serem potencialmente infinitas. Por outro lado, uma consequência dessa característica é que a distribuição dos eventos que compõem o conjunto de dados pode mudar ao longo do tempo, de modo que a estrutura deve adaptar-se à nova distribuição. Portanto, a mudança na distribuição dos dados faça com que os novos eventos incidam fora dos espaços de representação definidos inicialmente 4.1, será necessário readaptar a estrutura *STB-tree* para considerar essa nova região.

Nesse cenário, sempre que um novo evento é recebido pelo *STB-meter*, é necessário detectar se ele incide dentro ou fora dos espaços de representação da estrutura. Para tanto, é necessário verificar se: 1) os atributos espaciais do novo evento estão contidos no espaço representado pela estrutura de dados

Algoritmo 4.4 Procedimento *CalculaContagem*

Entrada: *Cell*: Célula da árvore de contagem.
S: *STB-tree*.
g: Granularidade desejada.
Saída: *SUM* Contagem.

1: $i \leftarrow 0$
2: $SUM \leftarrow 0$
3: $n_c atual \leftarrow$ período de contagem atual c_i mantido por S
4: $n_c total \leftarrow$ número de períodos de contagem total S
5: $C[\,] \leftarrow$ vetor que mantém os períodos de contagem mantidos por *Cell*
6: **para** $i = 0$ **até** g **faça**
7: $\quad SUM \leftarrow SUM + C[(n_c total - g + i) \bmod n_c total]$
8: **fim para**

hierárquica; 2) os atributos não espaciais do novo evento incidem dentro do espaço representado pela *box counting tree*. Caso o novo evento incida fora de alguma das representações, é realizada a expansão da *STB-tree*.

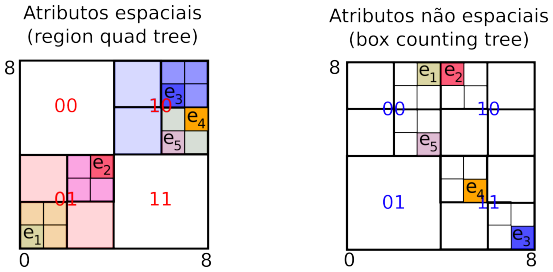

Figura 4.12: Espaços de representação após o processamento de todos os eventos da Tabela 4.1.

No exemplo anterior, após o processamento de todos os eventos presentes na Tabela 4.1, os espaços de representação mantidos pela *STB-tree* têm a forma apresentada na Figura 4.12. Quando um novo evento $e_6 = \{10, 10, 15, 15\}$ é recebido pelo *STB-meter*, separam-se seus atributos espaciais $a_s = \{10, 10\}$ e seus atributos não espaciais $a_v = \{15, 15\}$, e verifica-se se a_s está contido na representação espacial da *STB-tree* assim como se a_v está contido na *box-counting tree*. Na Figura 4.13 é possível verificar que o evento e_6 incide fora de ambas as representações, o que faz com que seja necessário expandir a *STB-tree* de modo que e_6 seja considerado na contagem de eventos utilizada no cálculo da dimensão fractal de correlação.

O processo de expansão consiste em criar um novo nível acima da estrutura *STB-tree*, ou seja, um novo nó para a estrutura da dados hierárquica e uma nova célula para a *Box Counting Tree*. Esse processo permite aumentar o espaço de representação sem a necessidade de realizar alterações na estrutura já existente e sem a necessidade de re-calcular as contagens de pontos, pois a estrutura já existente passa a ser referenciada pela nova célula da *box couting tree*, que por sua vez é referenciada pelo novo nó da estrutura de dados hierárquica.

Esse processo ocorre de acordo com a regra formação da estrutura de dados hierárquica. Desse modo, como a estrutura de dados utilizada neste trabalho é a *region quad tree*, essa expansão utiliza o lado r do espaço representado pela estrutura como referência para calcular o novo espaço que será representado.

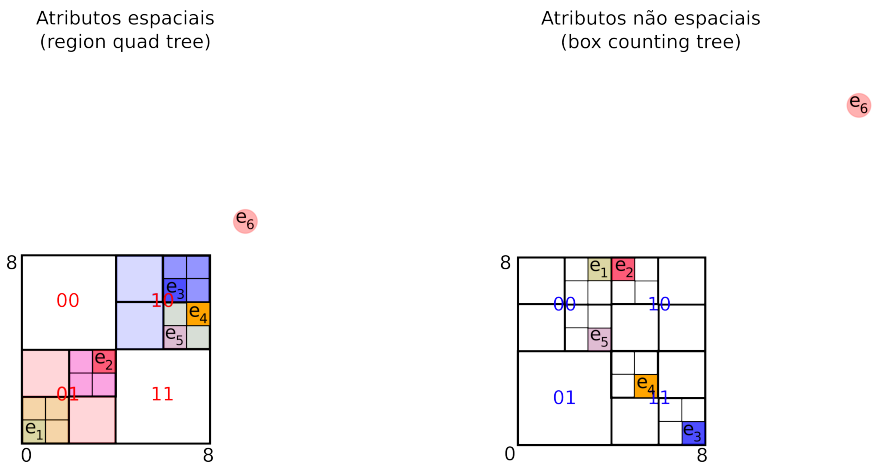

Figura 4.13: Evento e_6 que incide fora do espaço de representação da *STB-tree*.

Na Figura 4.14 é possível verificar a abordagem utilizada, onde inicialmente as dimensões do espaço de representação que deseja-se expandir é igual a r, portanto, para que seja possível manter o espaço anterior como um sub-nível da nova representação que será criada, deve-se ter um novo nó com dimensões iguais a $2r$.

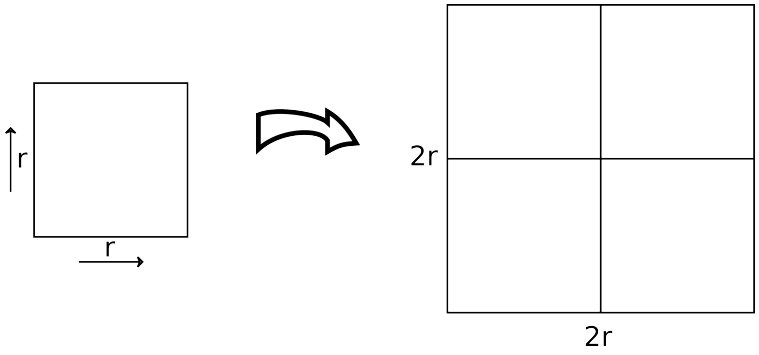

Figura 4.14: Processo de expansão de um nó da *quadtree* utilizado pela *STB-tree*.

Essa expansão ocorre no sentido do evento que extrapola a representação, ou seja, no momento de verificar se os atributos que compõem o evento estão contidos no espaço de representação mantido pela *STB-tree*, define-se em qual sentido essas estruturas devem se expandir de acordo com os limites (superior ou inferior) que o novo evento extrapola. Na Figura 4.15 é possível verificar que a expansão do espaço de representação ocorre sempre nas duas estruturas (*box counting tree* e estrutura de dados hierárquica). Essa abordagem é utilizada para que seja possível manter a coerência entre os níveis que constituem a *STB-tree*.

O processo de expansão da estrutura *STB-tree* é apresentado no Algoritmo 4.5 e ocorre do seguinte modo:

1. Inicialmente, verifica-se se o novo evento pode ser englobado pelos espaços mantidos pela estrutura *STB-tree*. Caso o novo evento incida fora da estrutura, o processo de expansão é iniciado.

2. Cria-se um novo nó da estrutura de dados hierárquica S_n. Esse novo nó deve englobar a região já

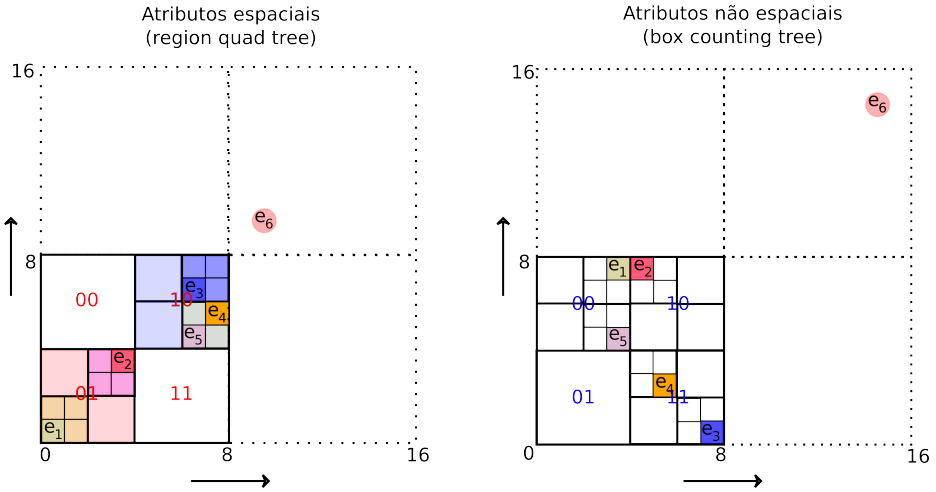

Figura 4.15: Expansão do espaço de representação da *STB-tree* no sentido do novo evento e_6.

Algoritmo 4.5 Procedimento *Expande*

Entrada: STB: *STB-tree*, e evento incidente fora da representação espacial
Saída: STB: *STB-tree*.

1: **enquanto** STB não englobar e **faça**
2: $S \leftarrow$ Estrutura Hierárquica utilizada por STB
3: $V \leftarrow$ *Box-couting tree* utilizada por STB
4: $S_n \leftarrow$ Nova instância da estrutura hierárquicas de dados com as dimensões de H expandidas no sentido dos atributos espaciais do evento e.
5: $V_n \leftarrow$ Nova instância da *Box-counting tree* com as dimensões de C expandidas no sentido dos atributos não espaciais do evento e.
6: $S_t \leftarrow$ Selecionar a sub-região de S_n que engloba o espaço de S.
7: Associa-se a S_t a estrutura de contagem V_n.
8: $V_t \leftarrow$ Selecionar a sub-região de V_n que engloba o espaço de V.
9: Associa-se a V_t a estrutura hierárquica de dados S.
10: Associa-se a STB as estruturas S_n e V_n como primeiro nível.
11: **fim enquanto**

representada pelos atributos espaciais do conjunto de dados além de expandir suas dimensões no sentido do novo evento que incide fora do espaço de representação.

3. Do mesmo modo que o item 2, cria-se uma célula da estrutura *box couting tree* V_n que englobe a região já representada pelos atributos não espaciais do conjunto de dados e que expande-se no sentido do novo evento que incide fora do espaço de representação.

4. O passo final é realizar as associações necessárias em S_n e V_n de modo que passem a referenciar a estrutura *STB-tree* já existente como uma célula da nova estrutura expandida, realizando a sumarização dos eventos incidentes em suas sub-regiões.

Um detalhe importante do Algoritmo 4.5 é a iteração realizada sobre o processo de expansão, composto pelo laço presente na linha 1 a 11. Isso ocorre pois a regra de expansão adotada pela estrutura de dados hierárquica, assim como a adotada pela *box counting tree*, pode fazer com que o espaço de representação expandido não englobe o novo evento em uma única iteração, sendo necessário realizar mais de um processo de expansão até que o novo evento seja englobado.

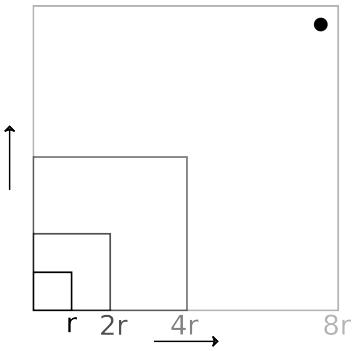

Figura 4.16: Processo iterativo de expansão do espaço de representação.

Na Figura 4.16 são ilustradas as iterações que ocorrem sobre o espaço de representação inicial com lado r, de modo a expandi-lo até que o novo evento seja englobado. Nesse exemplo, foram necessárias 3 iterações até que o espaço de representação seja capaz de englobar o novo evento, ou seja, o espaço inicial com lado r foi expandido para $2r$, $4r$ e por fim $8r$.

No exemplo, a *STB-tree* resultante do processo de expansão contendo o novo espaço de representação é apresentada na Figura 4.17. Pode-se verificar que dois novos níveis foram criados para representar a expansão realizada na estrutura de dados hierárquica assim como a expansão realizada sobre a *box counting tree*. Além disso, é possível notar como a estrutura original antes de ser expandida, passa a ser referenciada como um sub-nível da *STB-tree*.

Outra consequência da natureza evolutiva de *Data Streams* evolutivas é que a mudança na distribuição dos dados pode fazer com que regiões do espaço antes ocupadas deixem de serem utilizadas. Portanto, é necessário detectar situações em que os espaços representados pela *STB-tree* sejam maior que o necessário, visando liberar os recursos alocados desnecessariamente.

Esse processo ocorre no sentido oposto ao Algoritmo 4.5, ou seja, sempre que um período de contagem n_c é preenchido completamente, verifica-se se existe mais de uma sub-região preenchida na estrutura de dados hierárquica, assim como na *box counting tree*. Caso apenas um nó da estrutura de dados hierárquica e da *box counting tree* contenham dados no primeiro nível, a estrutura passa a considerar esse nó como nó raiz e descarta as demais regiões não utilizadas, liberando assim os recursos de memória não utilizados. Esse processo de contração do espaço de representação pode ser visto com maiores detalhes no Algoritmo 4.6.

Algoritmo 4.6 Procedimento *Contrai*

Entrada: *S*: *STB-tree*.
Saída: *S*: *STB-tree*.

1: **enquanto** Primeiro nível de estrutura hierárquica de *S* tiver apenas um nó e esse nó estiver associado a apenas uma célula da *Box-counting tree.* **faça**
2: $H \leftarrow$ Estrutura Hierárquica utilizada por *S*
3: $C \leftarrow$ *Box-couting tree* utilizada por *H*
4: Associa ao primeiro nível de *S* o próximo nível da estrutura de dados hierárquica associada a *C*.
5: Libera recursos alocados por *C*.
6: Libera recursos alocados por *H*.
7: **fim enquanto**

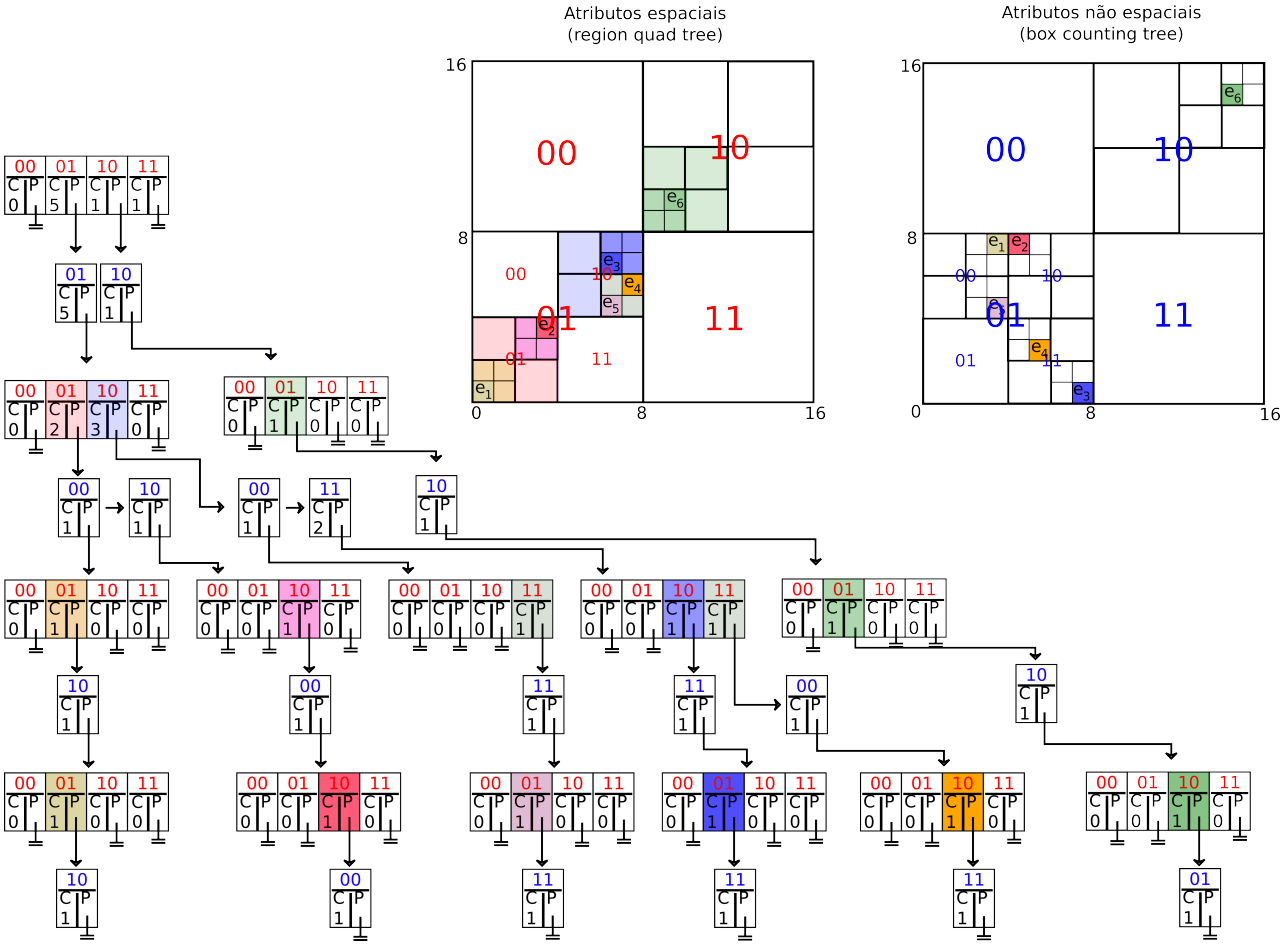

Figura 4.17: *STB-tree* resultante do processo de expansão.

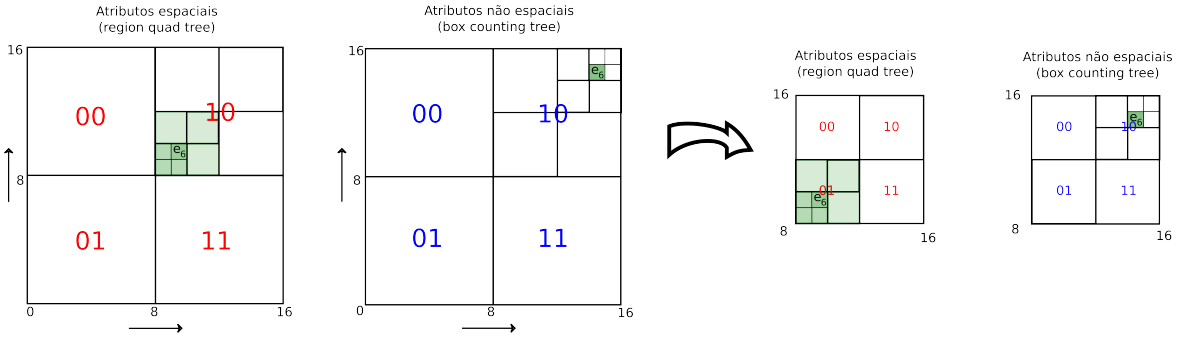

Figura 4.18: Processo de contração do espaço de representação.

Na Figura 4.18 é possível visualizar o processo de contração do espaço de representação. Supondo que a contagem dos 5 primeiros eventos foram descartados devido a movimentação da janela deslizante, o espaço de representação resultado é maior que o necessário para armazenar as contagens dos eventos atuais. Desse modo, seguindo o processo descrito pelo Algoritmo 4.6, a estrutura *STB-tree* sofre uma contração resultando na estrutura apresentada na Figura 4.19.

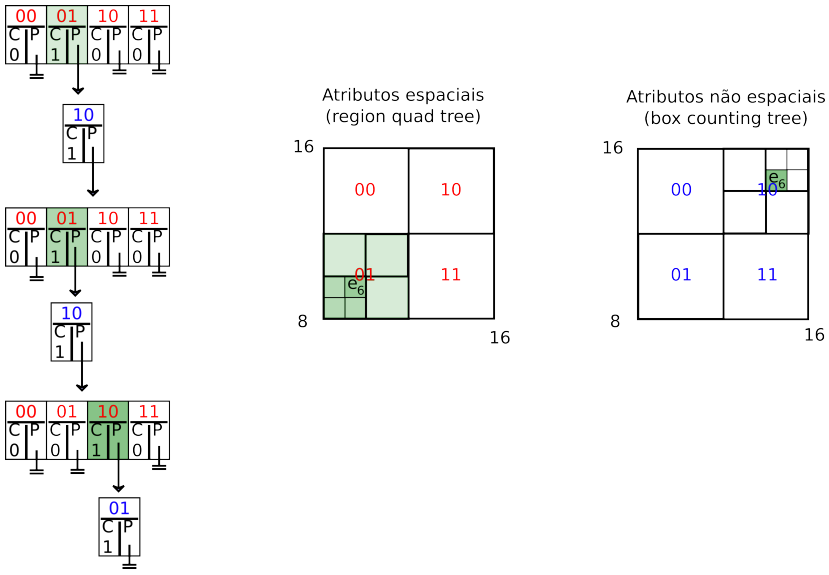

Figura 4.19: Estrutura *STB-tree* resultante do processo de contração.

4.3 Implementação no contexto da ferramenta ClimFractal Analyzer

O método *STB-meter*, descrito em detalhes na Seção 4.2, foi implementado no contexto da ferramenta *ClimFractal Analyzer* [Nunes 12]. Essa ferramenta tem como principal objetivo a análise de séries temporais climatológicas. Para tanto, diversos métodos de análise foram implementados na ferramenta como medidas estatísticas e a análise da variação da correlação dos atributos que compõem as séries temporais. A ferramenta *ClimFractal Analyzer* utiliza uma abordagem baseada em *workflows* para permitir a parametrização do fluxo da análise, de acordo com as necessidades do especialista do domínio.

Nesse cenário, a *ClimFractal Analyzer* considera uma série de construções individuais incorporadas em um cenário que permite a definição do fluxo de análise de forma dinâmica. Além disso, novas funcionalidades podem ser incorporadas, aumentando assim as possibilidades de análises realizadas pela ferramenta.

Na Figura 4.20 é possível visualizar a interface principal da ferramenta, em que à direita está localizada a interface que permite definir o fluxo dentre as funcionalidades disponíveis à esquerda. Essas funcionalidades estão dividias em 4 tipos principais:

1. **Fontes de Dados:** Tipos de dados que a ferramenta consegue utilizar em suas análises, como conexão com um Banco de Dados, arquivos no formato CSV, etc.

2. **Filtros:** Operação realizada sobre as Fontes de Dados para limitar o escopo de análise, como por exemplo restringir um intervalo temporal sobre as fontes de dados ou então um intervalo de latitude e longitude.

3. **Ferramentas de análise:** Operação realizada sobre as Fontes de Dados ou então sobre os Dados filtrados. Responsável por realizar determinada análise, como o cálculo de medidas estatísticas.

4. **Visualização:** Formato como dados e resultados são exibidos ao usuário. Por exemplo, um mapa com o conjunto de estações meteorológicas utilizadas na análise ou então séries temporais resultantes de uma análise temporal, como o acompanhamento da média de uma variável climatológica.

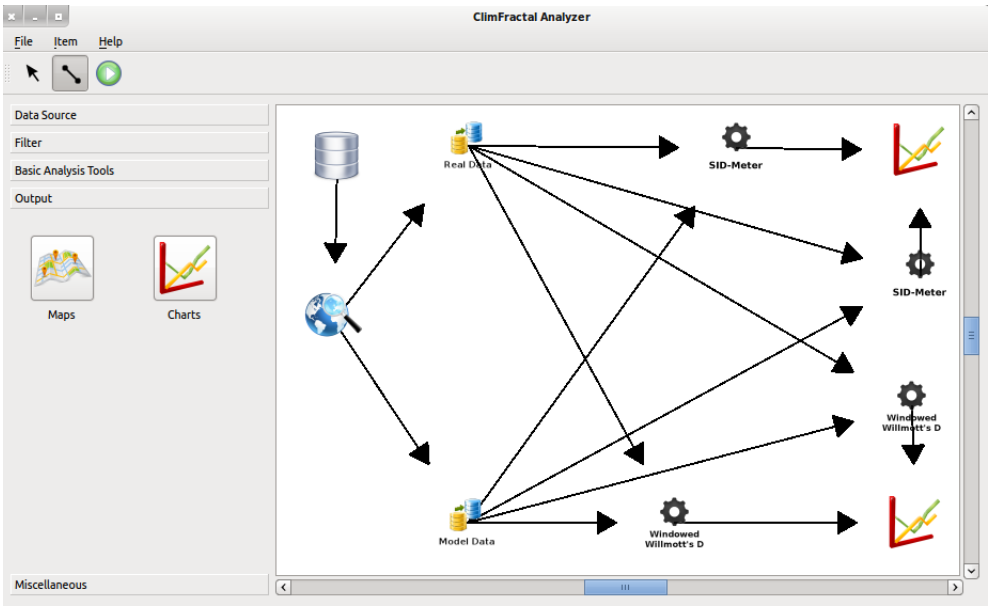

Figura 4.20: Tela principal da Ferramenta *ClimFractal Analyzer* exibindo uma configuração de um *workflow*.

Na Figura 4.21 é possível verificar um exemplo de *workflow* onde é realizado um filtro sobre para obter dados provenientes do Estado do Rio Grande do Sul (a esquerda) e exibindo a disposição da origem dos dados em um mapa (a direita).

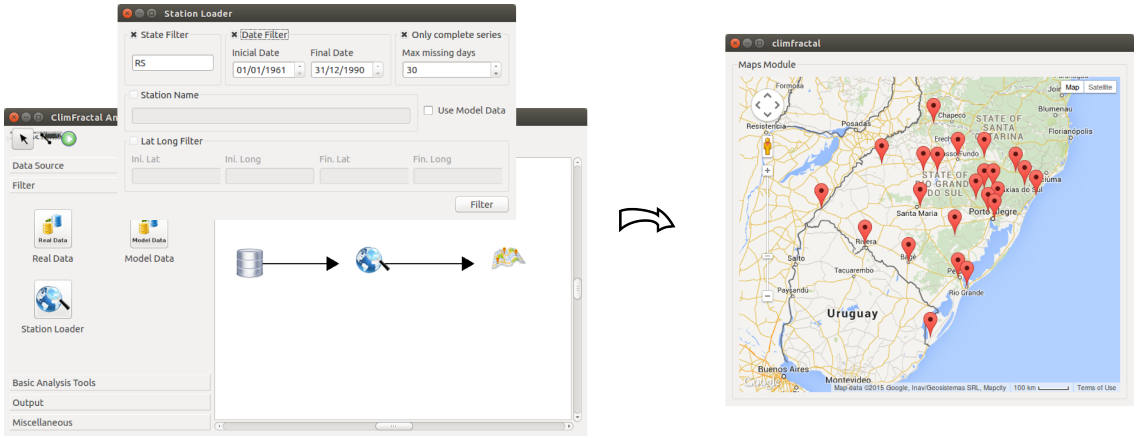

Figura 4.21: Exemplo de *worlkflow* que realiza o filtro de estações meteorológicas no estado do Rio Grande do Sul.

O método *STB-meter* foi incorporado à *ClimFractal Analyzer* como uma ferramenta de análise. Para tanto, os seguintes módulos foram criados:

1. ***Quadtree* :** A estrutura de dados hierárquica foi implementada proporcionando assim um mecanismo de representação espacial voltado para os conjuntos de dados meteorológicos utilizados pela Ferramenta *ClimFractal Analyzer*.

2. **Exibição da estrutura *Quadtree*:** O ambiente da *ClimFractal Analyzer* já proporciona um meio de visualização da disposição espacial das fontes de dados utilizadas, porém, foi necessário

estender essa visualização de modo que a estrutura de dados *Quadtree* seja exibida juntamente com as fontes de dados utilizadas.

3. **Janela deslizante estendida:** A janela deslizante estendida foi implementada de modo a permitir que determinar um escopo de análise considerando múltiplas granularidades temporais.

4. **STB-meter :** Por fim, o método *STB-meter* foi implementado de modo que seja possível realizar a análise espaço-temporal dos conjuntos de dados suportados pela Ferramenta *ClimFractal Analyzer*. A saída do método utiliza os métodos de visualização de resultados já disponíveis no ambiente da ferramenta.

Na Figura 4.22 são ilustradas as novas funcionalidades de representação da informação espacial dos conjuntos de dados utilizados no contexto da ferramenta *ClimFractal Analyzer*, de modo a utilizar os recursos já existentes no ambiente da ferramenta para representar a nova informação que é utilizada para montar a *Quadtree*.

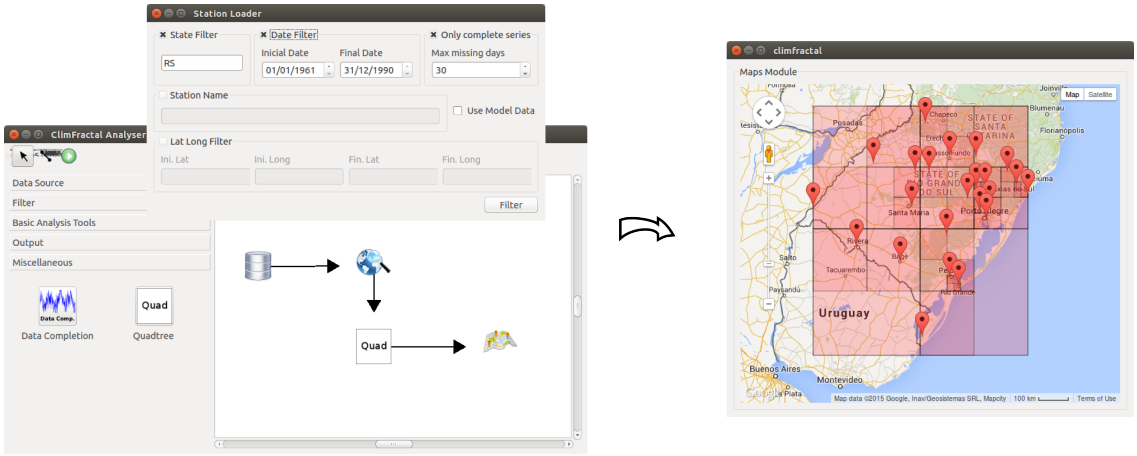

Figura 4.22: Exemplo de *worlkflow* que realiza o filtro de estações meteorológicas no estado do Rio Grande do Sul, porém, utilizando uma *quadtree* para realizar a representação da informação espacial.

Essa implementação inicial realizada no ambiente da ferramenta *ClimFractal Analyzer* permite realizar análises considerando aspectos temporais do conjunto de dados, por meio da utilização da abordagem de janelas deslizantes estendidas, que permitem uma multi-resolução temporal, além de também considerar aspectos espaciais, por meio da utilização da estrutura de dados hierárquicas.

4.4 Considerações Finais

Neste Capítulo foi apresentado o método *STB-meter*, que tem como principal objetivo permitir uma análise sobre a correlação de atributos que compõem *Data Streams* multidimensionais, considerando aspectos espaciais e não espaciais dos dados.

Para tanto, foram apresentadas as abordagens utilizadas na manipulação do *Data Stream* por meio de janelas deslizantes assim como a extensão realizada sobre essa abordagem de modo a permitir a utilização de diversas granularidades temporais simultaneamente. Além disso, essa abordagem permite o cálculo incremental da dimensão fractal de correlação D_2 sem a necessidade de recalcular as contagens mantidas pela *STB-tree* quando novos dados são recebidos.

Outro ponto relevante é a característica evolutiva presente em *Data Streams*, que tem como consequência a possibilidade de que a distribuição do conjunto de dados analisado mude ao longo do tempo. Desse modo, também foi apresentado o mecanismo de adaptação da estrutura interna *STB-tree*, utilizada para se adaptar a uma nova distribuição do conjunto de dados, ou seja, a *STB-tree* realiza uma evolução incremental de sua estrutura interna, sem a necessidade de reconstrução da estrutura inteira quando ocorre uma mudança na distribuição dos dados.

A abordagem conjunta utilizada pela estrutura que combina estrutura de dados hierárquica com a extensão do *box counting tree* e a extensão de janelas deslizantes permite que o *STB-meter* realize o cálculo de D_2 considerando multi-granularidades temporais e espaciais sem a necessidade de realizar a reconstrução da estrutura *STB-tree* para diferentes configurações de estudos experimentais.

Por fim, a Ferramenta *ClimFractal Analyzer* foi apresentada. Essa ferramenta foi estendida de modo a integrar o *STB-meter* como parte do seu conjunto de mecanismos de análise de dados. No Capítulo 5 são apresentados experimentos realizados utilizando a ferramenta com a implementação inicial do método *STB-meter*.

Capítulo 5

Experimentos

No Capítulo 4 foi apresentado o método *STB-meter* que tem como objetivo permitir a análise de *data streams* multidimensionais com atributos referentes a informações espaciais e não espaciais. Esse método teve sua implementação inicial realizada no ambiente da Ferramenta *ClimFractal Analyzer* de modo a utilizar uma série de recursos já disponíveis, como a obtenção do conjunto de dados analisados e também sua manipulação como *data streams*. Além disso, os novos recursos implementados permitem, por exemplo, a visualização da estrutura hierárquica de dados presente na *STB-tree*, de modo a permitir que o especialista do domínio da aplicação visualize a disposição espacial que o método utilizou para realizar a análise.

Nesse capítulo são apresentados experimentos realizados no ambiente dessa ferramenta para avaliação e validação do método proposto. Esses experimentos foram executados em um processador *Core i5 430M 2.27 Ghz* com 4GB de memória RAM, rodando o sistema operacional Ubuntu 12.04. Todos os experimentos utilizaram uma *STB-tree* configurada com $k = 20$, de modo que a estrutura tenha 20 níveis. Tendo em vista que cada nível é referente a um determinado refinamento do *Box Counting*, o *box count plot* terá 20 pontos, possibilitando detectar a curva e sua respectiva inclinação com maior precisão e resultando numa estimativa melhor de D_2.

Os experimentos executados têm como objetivo validar a eficácia do método *STB-meter*. Portanto, foram utilizados *data streams* sintéticos com comportamento previamente conhecido. De maneira semelhante, experimentos foram executados sobre dados reais, provenientes de estações meteorológicas de superfície e foram analisados por especialistas do domínio de modo a validar se os resultados obtidos estão de acordo com o esperado.

A Tabela 5.1 apresenta os conjuntos de dados utilizados nos experimentos, assim como uma breve descrição da origem desses pontos, número de eventos e também a dimensionalidade E dos conjuntos de dados.

Tabela 5.1: Data Streams.

Nome	Descrição	Número de Pontos	E
Dados Sintéticos	Dados Sintéticos - atributos com correlação conhecida	36000	6
Dados Reais SP	Dados Reais - Dados meteorológicos do Estado de São Paulo	262800	4
Dados Modelo ETA SP	Dados Simulados - Estado de São Paulo	262800	4

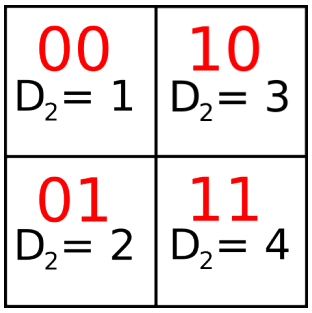

Figura 5.1: Correlação das sub-regiões do conjunto de dados sintético.

5.1 Experimentos com Dados Sintéticos

Para validar o método *STB-meter*, um conjunto de dados sintéticos foi construído de modo que cada sub-região do espaço seja composta por dados que apresentam um comportamento diferente. Esse conjunto de dados foi construído com o objetivo de verificar se o *STB-meter* calcula corretamente a dimensão fractal D_2 em diferentes sub-regiões do espaço. Logo, cada sub-região do espaço contém seus atributos não espaciais correlacionados de modo que sejam detectados diferentes valores de correlação em cada uma delas.

Outro objetivo do conjunto de dados sintéticos é verificar se a análise espacial ocorre adequadamente em múltiplos níveis, de modo que seja possível detectar comportamentos distintos do comportamento geral de uma região em suas sub-regiões. Por exemplo, dada uma região que apresenta baixa correlação em seus atributos, é possível que uma sub-região apresente um alto grau de correlação.

Com esse intuito, o conjunto de dados sintético foi construído do modo a dividir o espaço em 4 sub-regiões, onde cada sub-região apresenta um comportamento específico, definido por meio da correlação de seus atributos não espaciais. Na Figura 5.1 é apresentada a disposição do conjunto de dados formado por 2 atributos referente a informações espaciais e 4 atributos referente a informações não espaciais.

Esse conjunto de dados é formado por 72000 eventos dispostos da seguinte forma:

1. **Sub-região** 00 **:** Nessa sub-região, todos os atributos não espaciais estão correlacionados, e a dimensão fractal de correlação esperada é $D_2 = 1$.

2. **Sub-região** 01 **:** Essa sub-região é sub-dividida em duas sub-regiões, como apresentado na Figura 5.2, de modo que a sub-região com índice 00 contém pares de atributos não espaciais correlacionados, portanto, a dimensão fractal desse sub-conjunto é $D_2 = 2$. Por outro lado, a sub-região com

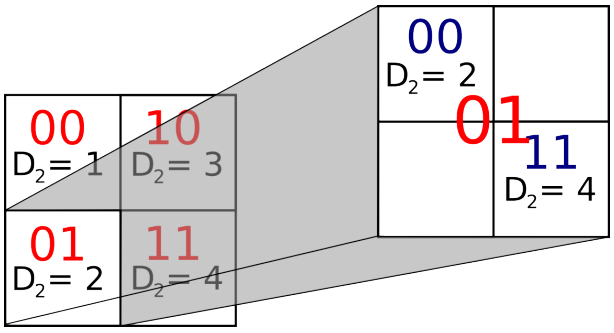

Figura 5.2: Correlação das sub-regiões da Região 01.

índice 11 é formada por atributos sem nenhuma correlação, de modo que a dimensão fractal seja $D_2 = 4$.

3. **Sub-região** 10 **:** Essa sub-região é composta por dois atributos não espaciais correlacionados e outros dois que não possuem correlação com nenhum outro atributo, de modo que a dimensão fractal dessa região seja $D_2 = 3$.

4. **Sub-região** 11 **:** Essa sub-região é composta por atributos não correlacionados, de modo que a dimensão fractal seja $D_2 = 4$.

Sobre esse conjunto de dados, o método *STB-meter* foi parametrizado para utilizar janelas deslizantes com períodos de contagem variando de 10 a 20, isto é, $n_c = 10$ e $r_c = 2$. Cada período de contagem foi configurado para conter 2400 eventos, ou seja, $n_e = 2400$ e $r_e = 1$. Essa configuração define uma janela deslizante com granularidades temporais 2400 e 4800. Na Figura 5.3 é possível verificar a dimensão fractal para cada sub-região que compõe o conjunto de dados sintético utilizando janelas de 2400 onde cada sub-região possui um comportamento distinto.

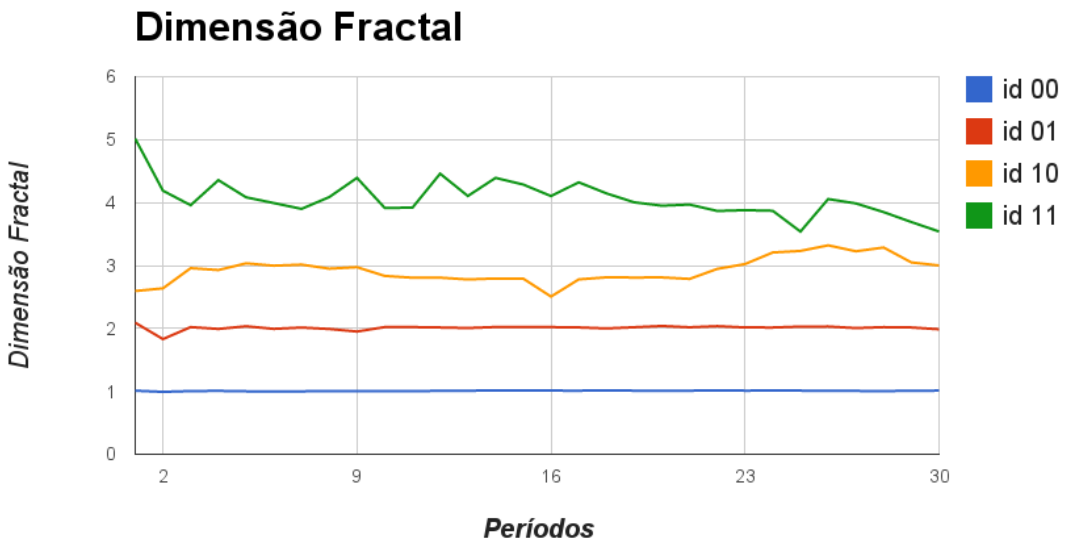

Figura 5.3: Dimensão Fractal da regiões do conjunto de dados sintéticos com $n_c = 10$ e $n_e = 2400$.

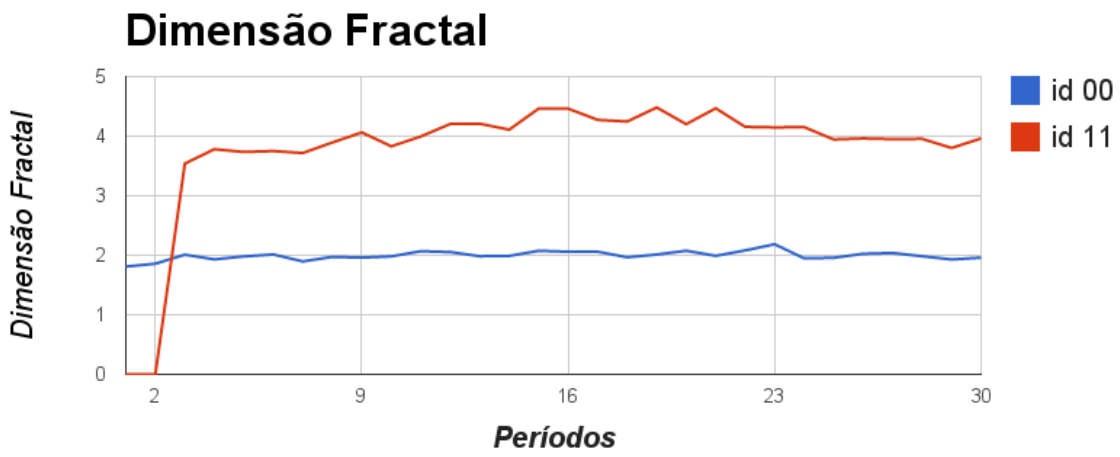

Figura 5.4: Dimensão Fractal da sub-região 01.

A dimensão fractal de cada sub-região, como pode ser visto na Figura 5.3, está de acordo com a correlação esperada apresentada na Figura 5.1. A sub-região 01, no entanto, é composta por outras duas sub-regiões com correlações conhecidas, como pode ser visto na Figura 5.2. Portanto, o cálculo da dimensão fractal foi realizado de modo a considerar apenas a sub-região 01 sobre a mesma estrutura montada para o conjunto de dados inteiro. Esse processo se dá através da seleção da seleção da região de interesse pelo usuário. Na Figura 5.4 é possível verificar que dimensão fractal para cada sub-região que compões a sub-região com índice 01, de modo a distinguir o comportamento do conjunto de dados em um sub-nível maior de refinamento, sendo possível verificar que a correlação detectada pelo método *STB-meter* está de acordo com o esperado, ou seja, em um segundo nível de refinamento sobre a sub-região 01 é foi possível verificar que a sub-região com índice 00 tem dimensão fractal $D_2 = 2$ e a sub-região com índice 11 tem $D_2 = 4$.

5.2 Experimentos com Dados Reais do Estado de São Paulo

Utilizando a ferramenta *ClimFractal Analyzer*, foram realizados experimentos com uma base de dados meteorológica com dados disponibilizados pelo Agritempo [1]. Essa base de dados é constituída por dados provenientes de estações meteorológicas e a localização das estações pode ser visualizada na Figura 5.5. Cada estação fornece medidas diárias de variáveis climatológicas.

O método *STB-meter* foi utilizado para analisar *data streams* criadas a partir de séries temporais climatológicas obtidas dessa rede de estações meteorológicas. Essas séries temporais são formadas por medidas diárias da temperatura média e precipitação.

O *data stream* construído a partir desse conjunto de dados utiliza informações provenientes da rede de estações meteorológicas para obter os atributos espaciais, assim como dados provenientes das séries temporais, utilizadas para obter os atributos não espaciais. A descrição dos atributos que compõe o *data stream* assim como a tipo de informação (espacial ou não espacial) podem ser visualizadas na Tabela 5.2

[1] http://www.agritempo.gov.br/

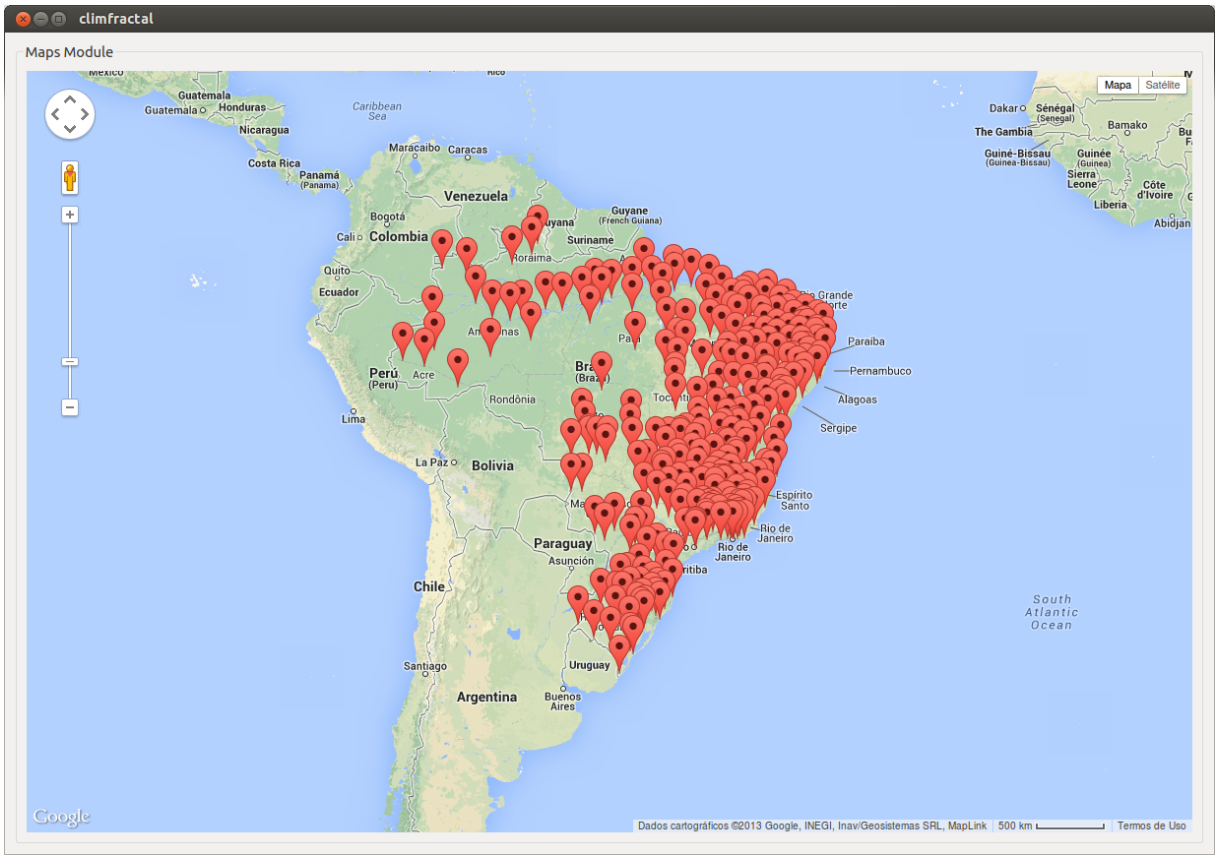

Figura 5.5: Disposição das estações meteorológicas disponibilizadas pelo Agritempo.

A construção do *data stream* ocorre de acordo com a ordem cronológica das séries temporais e com o conjunto de estações meteorológicas utilizados, ou seja, dado um conjunto de estações $S = \{s_1, ..., s_n\}$ e suas respectivas séries temporais $V = \{v_{s1}, ..., v_{sn}\}$ a partir de um momento t, o primeiro evento do *data stream* será composto pela latitude e longitude de s_1 e da precipitação e temperatura média de v_{s1} no instante t. Esse processo ocorre do mesmo modo para s_2 no instante t até chegar em s_n no instante t. O passo seguinte é realizar o mesmo processo utilizando o momento $t+1, t+2, ...$ até chegar no fim do conjunto de dados ou até um momento pré-determinado.

Uma característica importante presente na aquisição das séries de dados meteorológicas é que podem ocorrer erros no momento da aquisição, ou seja, podem existir valores faltantes ou medições inválidas na série temporal. Isso ocorre pode diversos motivos, como alguma falha nos sensores que realizam a medição ou até mesmo erro humano. Desse modo, o processo de construção do *data stream* deve considerar a possibilidade de valores faltantes ou incorretos. Nesse cenário, a Ferramenta *ClimFractal*

Tabela 5.2: Descrição dos atributos que compõem o *data stream*.

Índice do Atributo	Descrição	Classificação
1	Latitude	Espacial
2	Longitude	Espacial
3	Temperatura média	Não espacial
4	Precipitação	Não espacial

Analyzer permite que seja aplicado um filtro sobre as estações meteorológicas que serão utilizadas na construção do *data stream*. Na Figura 5.6 é apresentado o filtro realizado sobre as estações meteorológicas disponíveis na base de dados de modo a obter apenas as estações localizadas no estado de São Paulo, que contenham séries temporais climatológicas obtidas no intervalo de 1961 a 1990 e que, para cada ano, contenham no máximo 30 valores faltantes ou inválidos.

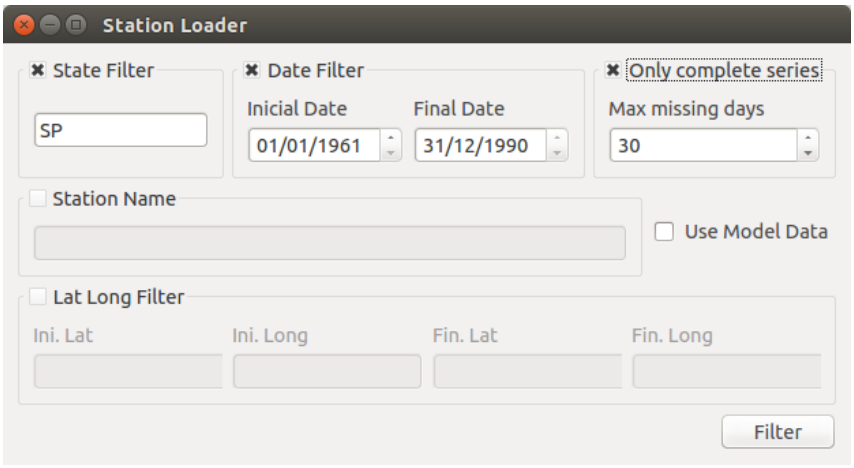

Figura 5.6: Filtro utilizado sobre as estações meteorológicas do estado de São Paulo.

O resultado do filtro apresentado na Figura 5.6 é um conjunto de dados formado por 24 estações meteorológicas, como pode ser visualizado na Figura 5.7, com séries temporais formadas por medidas diárias de temperatura média e precipitação no intervalo de 1961 a 1990.

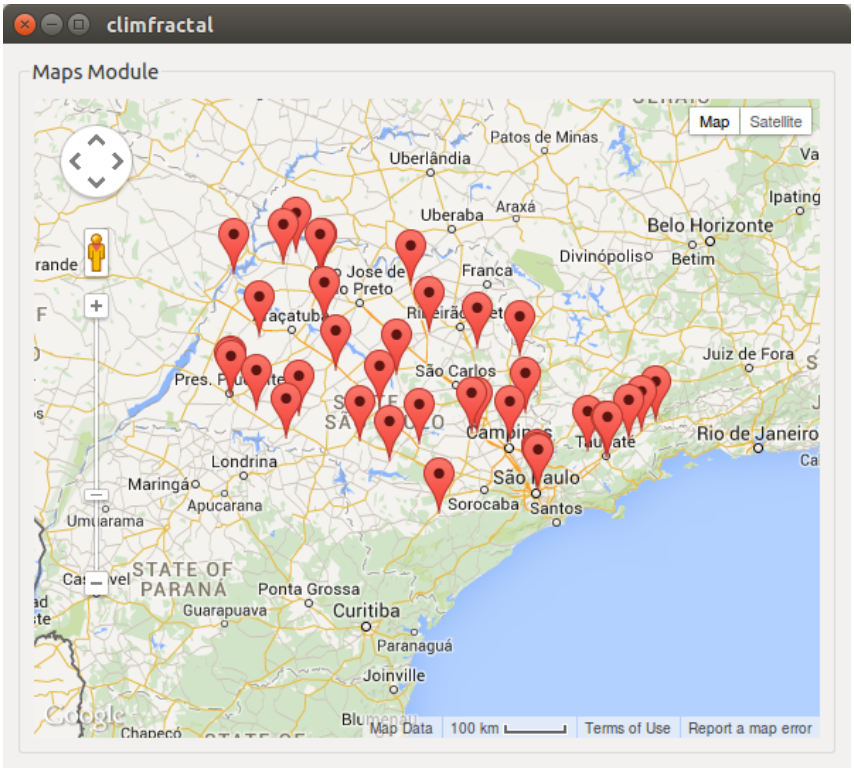

Figura 5.7: Estações meteorológicas resultantes do filtro apresentado na Figura 5.6.

Esse conjunto de dados é manipulado como um *data stream* multidimensional formado pelos atributos apresentados na Tabela 5.2. Sobre esse conjunto de dados foi aplicado o método *STB-meter* parametrizado de modo a utilizar uma abordagem temporal de 3 meses a 6 meses, ou seja, $n_c = 3$ e $r_c = 2$ com $n_e = 30$ e $r_e = 1$. Essas granularidades já se mostraram eficientes na análise da correlação de atributos em trabalhos anteriores, onde a análise da dimensão fractal permitiu detectar variações na correlação dos atributos em períodos referentes as eventos extremos como o *El Niño* e *Lá Niña* [Nunes 10].

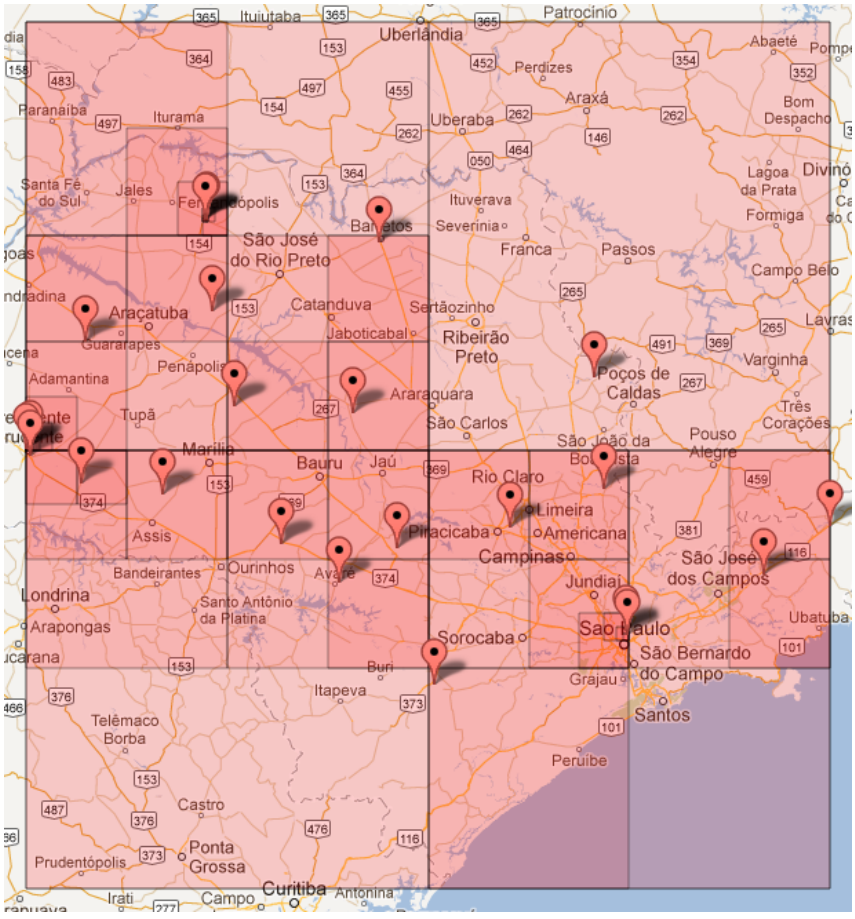

Figura 5.8: Estrutura hierárquica construída a partir da informação espacial do conjunto de dados.

As consultas realizadas para o cálculo da dimensão fractal abrangeram os três primeiros níveis da estrutura hierárquica, que pode ser verificada na Figura 5.8, ou seja, para os primeiros 3 níveis, todos os nós da estrutura de dados hierárquica em cada nível foram considerados individualmente para o cálculo da dimensão fractal.

A dimensão fractal, calculada utilizando granularidade temporal de 3 meses atualizada mensalmente, é apresentada na Figura 5.9 sendo que esse gráfico mostra a evolução da dimensão fractal D_2 pelo tempo, no intervalo de 1961 a 1990. É possível verificar os diferentes comportamentos para as diferentes granularidades espaciais obtidas a partir da estrutura de dados hierárquica. Pelo fato da *Quadtree* dividir cada região em 2^{N_S} sub-regiões, o resultado mostrado na Figura 5.9 apresenta um grande conjunto de

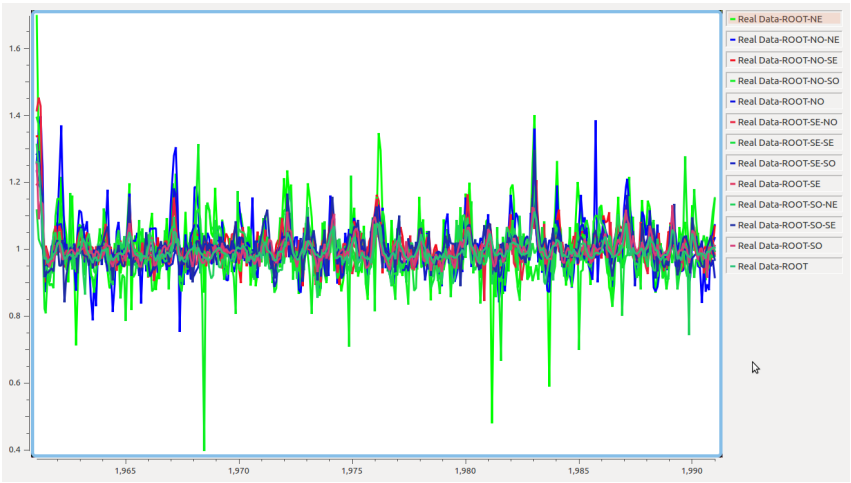

Figura 5.9: Cálculo da Dimensão Fractal D_2 utilizando 3 meses de granularidade temporal.

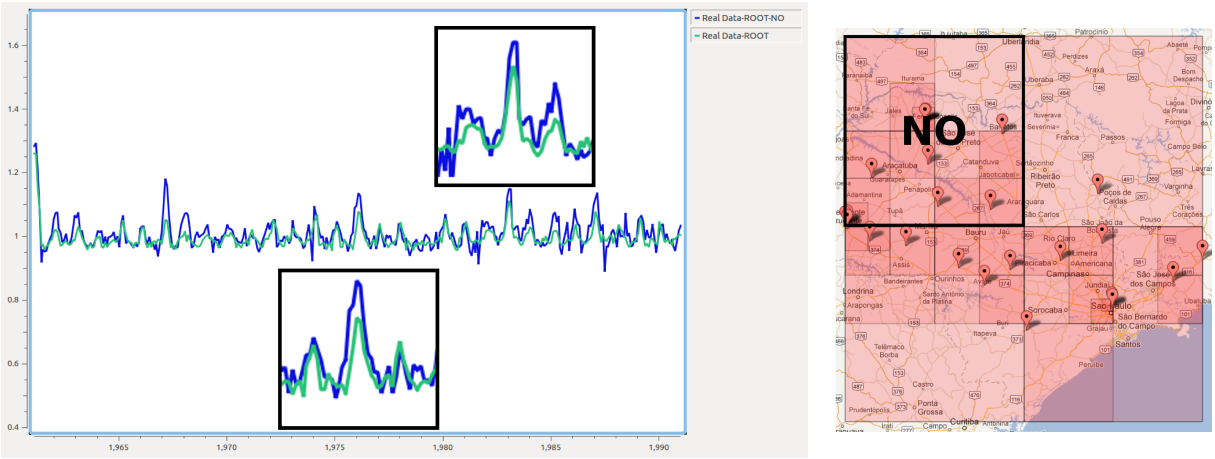

Figura 5.10: Destaque da Figura 5.9 para sub-região Noroeste do estado de São Paulo.

informações sobrepostas, o que dificulta a análise por parte do especialista do domínio da aplicação. Desse modo, a Ferramenta *ClimFractal Analyzer* permite manipular a apresentação dos resultados visando facilitar a visualização e interpretação dos resultados.

Nas Figura 5.10, 5.11, 5.12 é apresentado o comportamento individual de D_2 para as sub-regiões Noroeste, Sudeste e Sudoeste, respectivamente, obtido a partir da Figura 5.9. Além disso, o comportamento de todo o conjunto de dados é exibido simultaneamente, permitindo realizar a comparação entre o comportamento geral dos dados com o comportamento das sub-regiões. A sub-região Nordeste não foi inclusa na análise pois há apenas uma estação meteorológica localizada nessa sub-região.

Nos anos de 82/83, o evento *El Niño* ocorreu com maior intensidade, esse evento pode fazer com que a correlação entre as variáveis climáticas mude, já que influencia no regime de chuva. No ano de 76 também ouve a ocorrência desse evento com bastante intensidade. A dimensão fractal calculada utilizando o método *STB-meter* detectou a variação na correlação dos atributos temperatura média e precipitação nesses períodos.

O Sudeste do Brasil é uma região de transição entre efeitos opostos de eventos *El Niño* e possui uma topografia bastante acidentada o que pode responder por grande variabilidade espacial dos efeitos desses eventos. Analisando-se as diferentes regiões observa-se que os eventos de *El Niño* influencia o regime

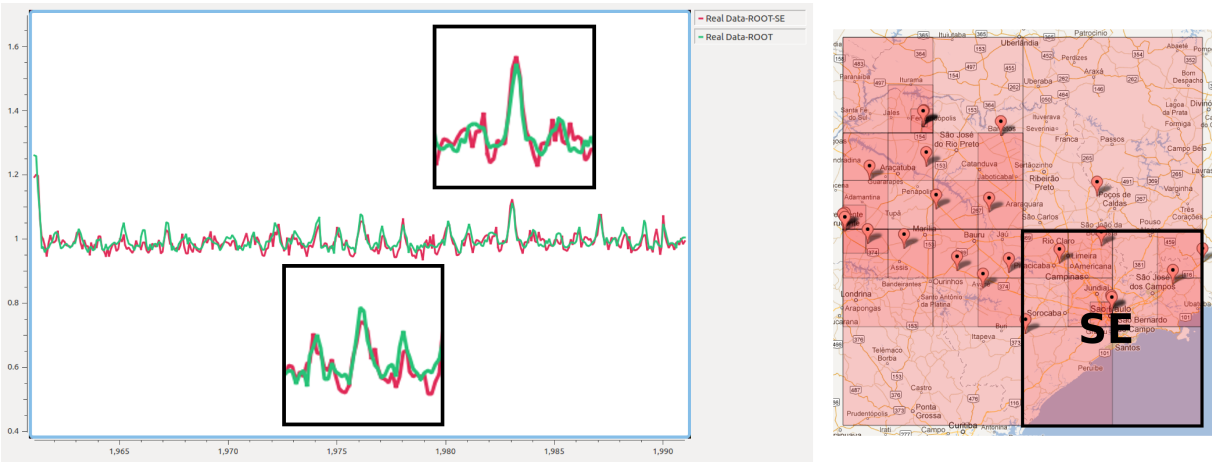

Figura 5.11: Destaque da Figura 5.9 para sub-região Sudeste do estado de São Paulo.

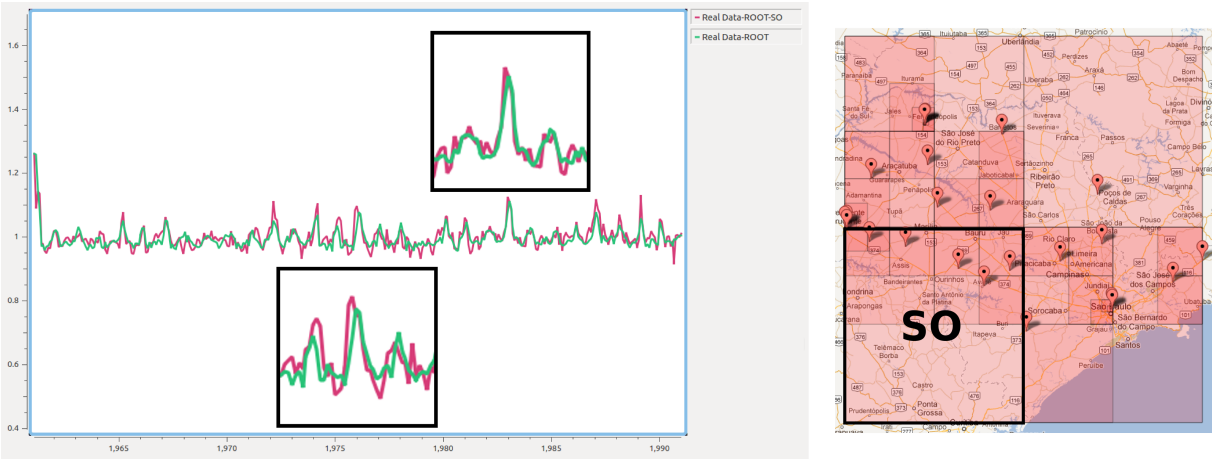

Figura 5.12: Destaque da Figura 5.9 para sub-região Sudoeste do estado de São Paulo.

de chuva no Noroeste do Estado de forma mais acentuada do que nas demais regiões do Estado.

Com a análise espacial realizada utilizando o método *STB-meter*, foi possível detectar como cada região apresenta um comportamento distinto diante do eventos extremos como *El Niño*. Lembrando que todas as análises foram realizadas sobre a mesma estrutura de dados, sem a necessidade de realizar novas leituras dos dados.

5.3 Experimentos com Dados Simulados do Estado de São Paulo

Em outro experimento, foram utilizados dados provenientes do modelo climático ETA, desenvolvido pela Universidade de Belgrado e o Instituto Hidrometeorológico da antiga Iugoslávia, posteriormente operacionalizados pelo *Nacional Centers for Environmental Prediction* (NCEP) [Black 94], instalado no Centro de Previsão de Tempo e Estudos Climáticos (CPTEC) em 1996 [Chou 96]. Esse experimento tem como objetivo permitir a comparação entre o comportamento dos conjuntos de dados reais e dos dados simulados fornecidos pelo CPTEC, de modo a verificar se a correlação entre os atributos utilizados nos modelos climáticos está em conformidade com a apresentada nos dados reais.

Esse modelo simula uma série de variáveis meteorológicas, como a precipitação, temperatura, umidade do ar e pressão atmosférica sobre uma grade de pontos com variação de 0.4×0.4 de latitude e

longitude sobre toda a América do Sul e uma pequena região da América Central, com valores diários para cada variável. De modo a viabilizar a comparação entre os conjuntos de dados reais e de dados gerados pelo modelo climático, foram selecionados os pontos mais próximos de cada estação meteorológica utilizada no experimento descrito na Seção 5.2. Além disso, o conjunto de atributos selecionados para realizar a análise são os mesmo presentes nos dados reais, de modo que o resultado da correlação mantenha a mesma semântica em ambos os experimentos. O intervalo temporal do conjunto de dados simulados é o mesmo utilizado para o conjunto de dados reais, ou seja, de 1961 a 1990.

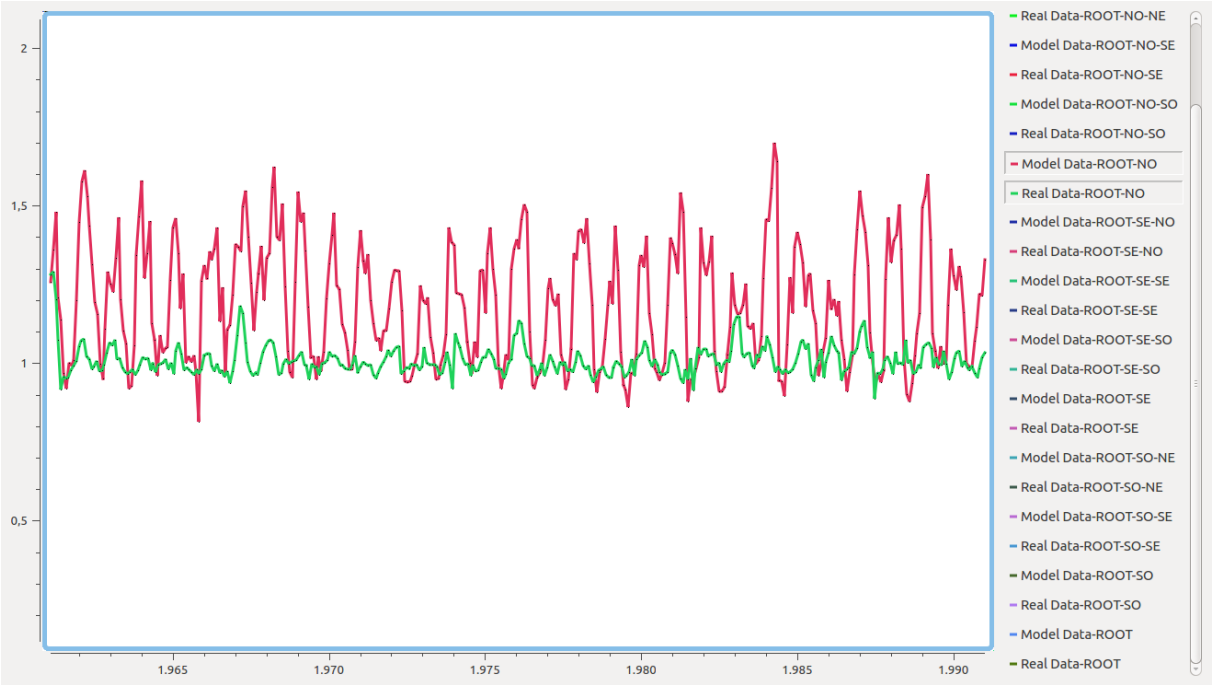

Figura 5.13: Comparação da correlação entre dados reais e dados simulados para a região Noroeste do Estado de São Paulo .

Nas Figuras 5.13, 5.14 e 5.15 é apresentado o resultado a variação da correlação para cada sub-região apresentada na Figura 5.8 para os dados provenientes do modelo ETA. Em conjunto com essas informações, também é apresentada a correlação dos atributos para os dados reais.

Em todas as figuras, é possível verificar que a correlação dos atributos simulados é menor que a dos atributos reais, ou seja, o valor da dimensão fractal de correlação para o conjunto de dados simulado é, em geral, maior que para os dados reais. Essa análise permite diferenciar os dados reais dos dados provenientes do modelos climático, pois a correlação entre as variáveis climáticas identificada nos dados reais são consideravelmente diferente das obtidas nos dados simulados. Esses resultados sugerem que existe espaço para melhorias no modelo climático e que a análise espaço-temporal realizada por meio do método *STB-meter* pode contribuir nessa tarefa.

Essas resultados são relevantes na detecção de mudanças climáticas e de eventos extremos, além disso, esses resultados podem ser úteis na agrometeorologia pois diversas culturas agrícolas são extremamente vulneráveis a variações climáticas. Também é possível utilizar essa análise como medida de conformidade em modelos climáticos, de modo a verificar se a correlação entre os atributos simulados

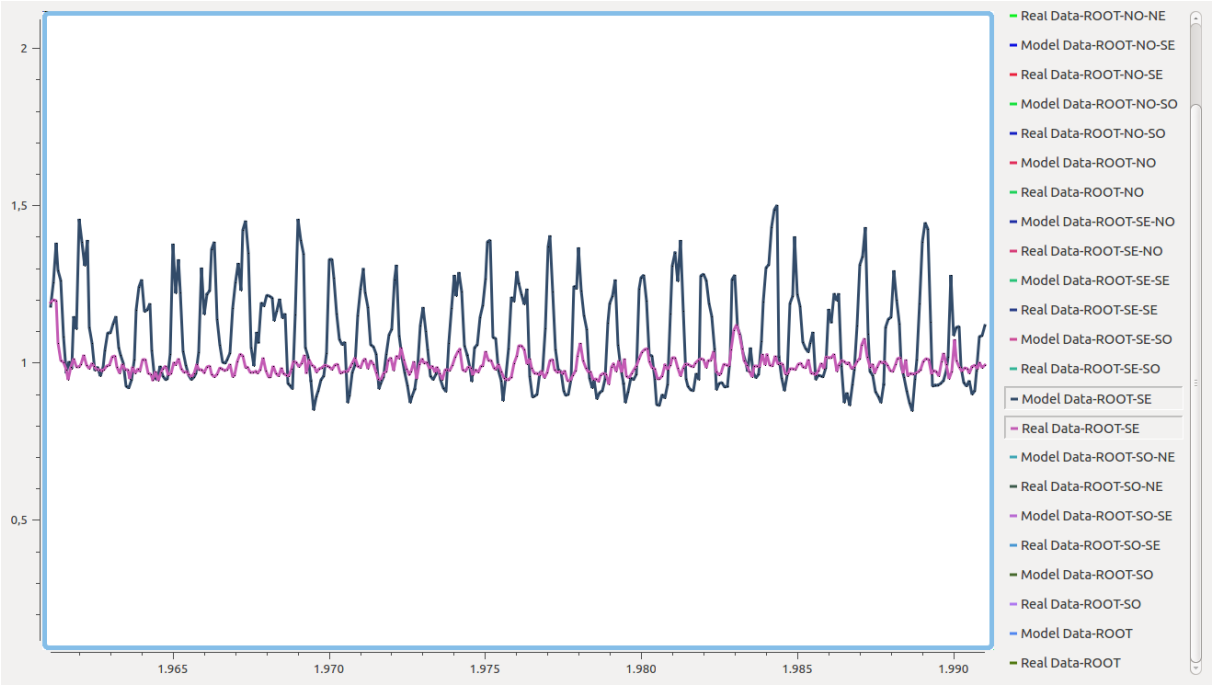

Figura 5.14: Comparação da correlação entre dados reais e dados simulados para a região Sudeste do Estado de São Paulo .

estão de acordo com o esperado.

5.4 Conclusões

Neste Capítulo uma série de experimentos foram apresentados utilizando conjuntos de dados sintéticos e reais. Os experimentos com dados sintéticos têm como objetivo verificar se o método *STB-meter* é capaz de detectar a variação da correlação dos atributos que compõem o conjunto de dados ao longo do tempo em diversas granularidades espaciais. Em outra etapa, experimentos foram realizados sobre dados meteorológicos reais, com o acompanhamento de especialistas do domínio da aplicação, de modo a verificar se os resultados estão de acordo com as expectativa dos especialistas. No Capítulo 6 são apresentados as conclusões e as publicações resultantes deste trabalho.

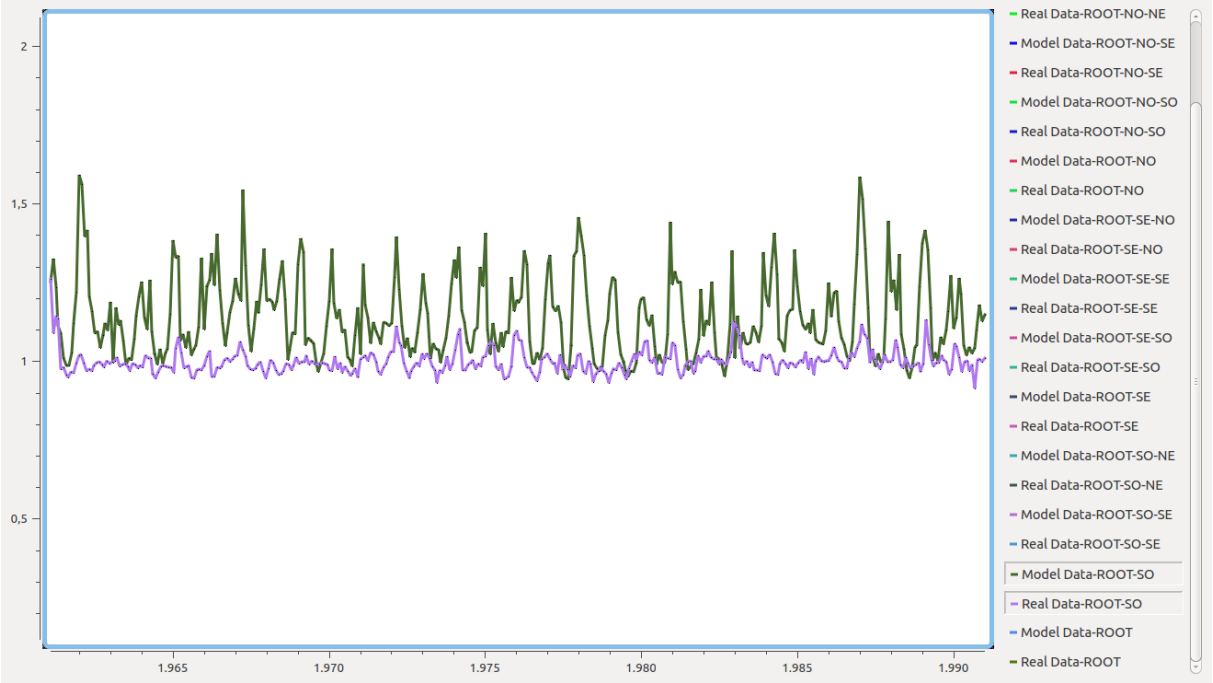

Figura 5.15: Comparação da correlação entre dados reais e dados simulados para a região Sudoeste do Estado de São Paulo .

Capítulo 6

Conclusões

Neste trabalho de mestrado foi proposto e desenvolvido um método de análise de *Data Streams* multidimensionais. Esses dados, em geral, são caracterizados por grandes volumes de dados gerados por processos síncronos ou assíncronos e podem ser potencialmente infinitos. Essa característica traz algumas limitações, como a impossibilidade de persistir esses dados e a impossibilidade de realizar mais que uma leitura de cada evento que constitui o *data stream*, que são impeditivas para os Gerenciadores de Bancos de Dados atuais, assim como diversas técnicas de mineração de dados disponíveis.

Além disso, *Data Streams* podem conter atributos que são referentes a informações espaciais e não espaciais. Essas informações podem ser utilizadas para realizar uma análise que considere os aspectos espaciais do conjunto de dados assim como a evolução dos aspectos não espaciais de modo a permitir uma análise conjunta desses dois aspectos em uma abordagem espaço temporal.

Nesse cenário, este trabalho propõe o método *STB-meter*, apresentado com maiores detalhes no Capítulo 4. Esse método tem como objetivo permitir a análise de *Data Streams* multidimensionais que tenham informações espaciais e não espaciais, de modo a viabilizar uma análise da correlação dos atributos não espaciais ao longo do tempo, considerando aspectos espaciais do conjunto de dados.

6.1 Contribuições

Nesta dissertação de mestrado foi apresentado o método *STB-meter*, que tem como objetivo permitir análises de *data streams* multidimensionais em múltiplas resoluções espaciais e temporais de forma simultânea. Esse método utiliza uma abordagem baseada na construção da estrutura *STB-tree*, que realiza a integração de características presentes na estrutura *box-counting tree*, estruturas de dados hierárquicas e também de janelas deslizantes.

As principais características do *STB-meter* são :

1. Cálculo incremental de D_2, não sendo necessário recalcular as contagens quando novos eventos são recebidos e processados pelo método.

2. Evolução incremental da estrutura *STB-tree*, sem a necessidade de reconstruir a estrutura quando ocorre mudanças na distribuição dos dados do *stream* ao longo do tempo.

3. Permitir múltiplas resoluções espaciais e temporais utilizando uma única estrutura.

O Capítulo 4 apresenta os algoritmos utilizados pelo método assim como alguns exemplos da construção e manutenção da estrutura interna *STB-tree* utilizada pelo método. Além disso, o método foi implementado na Ferramenta *ClimFractal Analyzer* como um recurso capaz de aumentar as possibilidades de análise realizadas por especialistas.

No Capítulo 5 é realizada uma série de experimentos utilizando conjuntos de dados sintéticos com comportamento espacial e também da correlação de seus atributos previamente conhecidos, dados reais proveniente de estações meteorológicas de superfície e também dados provenientes de modelos climáticos. O experimento sobre dados sintético foi possível verificar se o método *STB-meter* calcula corretamente a correlação D_2 em diversas granularidades espaciais. No experimento sobre dados meteorológicos reais foi possível detectar a mudanças na correlação dos atributos que compõem a *stream* em períodos de tempo referentes a eventos extremos e, além disso, foi possível detectar como cada sub-região espacial dos dados apresenta variações distintas diante do mesmo evento e, por fim, no experimento sobre dados provenientes de modelo climático, foi possível distinguir claramente a correlação entre os atributos dos dados reais em comparação com a dos dados de modelo climático em todas as diferentes sub-regiões dos dados. Esses resultados sugerem que ainda existe espaço para melhoria do modelo climático, de modo a considerar a correlação entre os atributos e, além disso, a análise por meio do método *STB-meter* pode contribuir nessa tarefa.

6.2 Publicações

Ao longo das etapas de desenvolvimento deste trabalho de mestrado, os resultados obtidos foram publicados em quatro artigos, como segue:

- Diversas funcionalidades foram acrescidas no ambiente da Ferramenta *ClimFractal Analyzer*, como apresentado na Seção 4.3. Essas funcionalidades juntamente com alguns recursos anteriores foram apresentados na seguinte publicação:

  NUNES, S. A.; COLIBARDINI, J. E. M.; COLTRI, P. P.; AVILA, A. M. H.; ROMANI, L. A. S.; TRAINA JUNIOR, C.; TRAINA, A. J. M.; SOUSA, E. P. M. ClimFractal Analyser: um ambiente de análise de séries temporais climáticas baseado em workflows. In: 27th Brazilian Symposium on Databases - Demos and Applications Session, 2012, São Paulo. Proceedings of the 27th Brazilian Symposium on Databases - 9th Demos and Applications Session, p. 25–30, 2012.

- Em estudos experimentais foram realizadas análises sobre dados provenientes de modelos climáticos a fim de detectar a conformidade das correlações dos atributos desses dados quando comparados a dados reais.

  Esses resultados resultaram nas seguintes publicações:

  NUNES, S. A.; ROMANI, L. A. S.; COLTRI, P. P.; AVILA, A. M. H.; TRAINA, A. J. M.; SOUSA, E. P. M. To be or not to be: fractal analysis of data streams from a regional climate change model. In: *27th Symposium On Applied Computing (SAC 2012)*, Trento, Itália, p. 831–832, 2012.

NUNES, S. A. ; ROMANI, L. A. S. ; AVILA, A. M. H. ; COLTRI, P. P. ; TRAINA JUNIOR, C. ; CORDEIRO, R. L. F. ; TRAINA, A. J. M. ; SOUSA, E. P. M. Analysis of Large Scale Climate Data: How Well Climate Change Models and Data from Real Sensor Networks Agree?. In: *The 2nd International Workshop on Large Scale Network Analysis* (em conjunto com *22nd International World Wide Web Conference - WWW 2013*), 2013, Rio de Janeiro. WWW 2013 Companion Publication. Rio de Janeiro. p. 517–526, 2013.

- Por fim, o método *STB-meter* proposto neste trabalho foi apresentado na publicação a seguir, onde foram apresentadas as principais características e técnicas utilizadas no desenvolvimento do método.

  NUNES, S. A. ; ROMAIN, L. A. S. ; AVILA, A. M. H.; COLTRI, P. P. ; TRAINA, A. J. M. ; SOUSA, E. P. M. Finding Spatio-Temporal Patterns In: *Multidimensional Data Streams. Journal of Information and Data Management - JIDM*, v. 4, p. 327-340, 2013.

6.3 Considerações Finais e Trabalhos Futuros

Neste trabalho de mestrado foi apresentado o método *STB-tree* e sua implementação inicial que utiliza a *Region Quadtree* como estrutura hierárquica para realizar a representação dos atributos espaciais de *Data Stream* multidimensionais. A *Region Quadtree* realiza a divisão do espaço de representação de forma recursiva e em partes iguais, sendo assim muito útil na análise meteorológica realizada neste trabalho, pois não é necessário se preocupar com limites territoriais.

Contudo, em uma análise que considere outros aspectos da representação espacial como por exemplo a divisão do espaço considerando fronteiras geográficas, topológicas, etc, a utilização da *Region Quadtree* pode não ser a ideal. Nesse cenário, a implementação de outras estruturas hierárquicas de dados é tida como uma funcionalidade importante a ser desenvolvida futuramente, de modo a aumentar as possibilidades de análises realizadas pelo método *STB-meter*.

Além disso, outros domínios de aplicação onde os dados que se deseja analisar podem ser manipulados como *Data Streams* multidimensionais podem ser utilizados no processo de análise, de modo a verificar a usabilidade do método em outros cenários. Por fim, tendo em vista o grande número de sub-regiões e granularidades temporais geradas pelo método, técnicas de detecção de anomalias poderiam se empregadas, permitindo a detecção automática de anomalias pelo método.

Referências Bibliográficas

[Aggarwal 03] Charu C. Aggarwal. *A framework for diagnosing changes in evolving data streams*. In Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data (SIGMOD'03), pages 575–586. ACM Press, June 2003.

[Aggarwal 04] Charu C. Aggarwal, Jiawei Han, Jianyong Wang & Philip S. Yu. *A Framework for Projected Clustering of High Dimensional Data Streams.* In Proceedings of the Thirtieth International Conference on Very Large Data Bases (VLDB'04), pages 852–863, Toronto, Canada, August/September 2004. Morgan Kaufmann.

[Aggarwal 07] Charu C. Aggarwal, editeur. Data streams - models and algorithms, volume 31 of *Advances in Database Systems*. Springer, 2007.

[Aggarwal 14] Charu C. Aggarwal. *A Survey of Stream Classification Algorithms*. In Data Classification: Algorithms and Applications, pages 245–274. 2014.

[Arantes 05] Adriano Siqueira Arantes. *Consultas por Similaridade Complexas em Gerenciadores Relacionais*. PhD thesis, Instituto de Ciências Matemáticas e de Computação - ICMC, Universidade de São Paulo - USP, São Carlos, São Paulo, 2005.

[Baioco 07] Gisele Busichia Baioco, Agma Juci Machado Traina & Caetano Traina. *MAMCost: Global and Local Estimates leading to Robust Cost Estimation of Similarity Queries*. In Proceedings of the 19th International Conference on Scientific and Statistical Database Management (SSDBM 2007), pages 6–16, Banff, Canada, 2007. ACM Press.

[Barbará 00] Daniel Barbará & Ping Chen. *Using the Fractal Dimension to Cluster Datasets*. In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'00), pages 260–264, Boston, USA, August 2000.

[Belussi 95] Alberto Belussi & Christos Faloutsos. *Estimating the Selectivity of Spatial Queries using the 'Correlation' Fractal Dimension*. In Proceedings of the 21th International Conference on Very Large Data Bases (VLDB'95), pages 299–310, Zurich, Switzerland, September 1995.

[Belussi 98] Alberto Belussi & Christos Faloutsos. *Self-Spatial Join Selectivity Estimation Using Fractal Concepts*. ACM Transactions on Information Systems, vol. 16, no. 2, pages 161–201, April 1998.

[Black 94] Thomas L. Black. *The New NMC Mesoscale Eta Model: Description and Forecast Examples*. Wea. Forecasting, vol. 9, no. 2, pages 265–278, June 1994.

[Böhm 00a] Christian Böhm. *A Cost Model for Query Processing in High Dimensional Data Spaces*. ACM Transactions on Database Systems, vol. 25, no. 2, pages 129–178, June 2000.

[Böhm 00b] Christian Böhm & Hans-Peter Kriegel. *Dynamically Optimizing High-Dimensional Index Structures*. In Advances in Database Technology - EDBT 2000, 7th International Conference on Extending Database Technology, volume 1777 of *Lecture Notes in Computer Science*, pages 36–50, Konstanz, Germany, March 2000. Springer.

[Chakrabarti 02] Kaushik Chakrabarti, Eamonn Keogh, Sharad Mehrotra & Michael Pazzani. *Locally adaptive dimensionality reduction for indexing large time series databases*. ACM Transactions on Database Systems (TODS), vol. 27, no. 2, pages 188–228, June 2002.

[Chaudhry 05] Nauman A. Chaudhry. *Introduction to Stream Data Management*. In Nauman A. Chaudhry, Kevin Shaw & Mahdi Abdelguerfi, editeurs, Stream Data Management, volume 30 of *Advances in Database Systems*, pages 1–13. Springer, 2005.

[Chou 96] S.C Chou. *Modelo regional Eta. Climanálise*. Rapport technique, Instituto Nacional de Pesquisas Espaciais, 1996.

[Cordeiro 10] Robson L. F. Cordeiro, Agma Juci Machado Traina, Christos Faloutsos & Caetano Traina. *Finding Clusters in subspaces of very large, multi-dimensional datasets*. In Proceedings of the 26th International Conference on Data Engineering (ICDE 2010), pages 625–636, Long Beach, California, USA, 2010. IEEE.

[Faghmous 12] James H. Faghmous, Yashu Chamber, Shyam Boriah, Frode Vikebø, Stefan Liess, Michel dos Santos Mesquita & Vipin Kumar. *A Novel and Scalable Spatio-Temporal Technique for Ocean Eddy Monitoring*. In Jörg Hoffmann & Bart Selman, editeurs, Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, July 22-26, 2012, Toronto, Ontario, Canada. AAAI Press, 2012.

[Faghmous 14] JamesH. Faghmous & Vipin Kumar. *Spatio-temporal Data Mining for Climate Data: Advances, Challenges, and Opportunities*. In Wesley W. Chu, editeur, Data Mining and Knowledge Discovery for Big Data, volume 1 of *Studies in Big Data*, pages 83–116. Springer Berlin Heidelberg, 2014.

[Faloutsos 94] Christos Faloutsos & Ibrahim Kamel. *Beyond Uniformity and Independence: Analysis of* R*-trees Using the Concept of Fractal Dimension*. In Proceedings of the 13th ACM Symposium on Principles of Database Systems (PODS'94), pages 4–13, Minneapolis, USA, May 1994. ACM Press.

[Faloutsos 96] Christos Faloutsos. Searching multimedia databases by content. Kluwer Academic Publishers, Norwell, USA, 1996.

[Faloutsos 00] Christos Faloutsos, Bernhard Seeger, Agma Juci Machado Traina & Caetano Traina. *Spatial Join Selectivity Using Power Laws*. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data (SIGMOD'00), pages 177–188, Dallas, USA, May 2000.

[Gaber 05] Mohamed Medhat Gaber, Arkady B. Zaslavsky & Shonali Krishnaswamy. *Mining data streams: a review*. SIGMOD Record, vol. 34, no. 2, pages 18–26, 2005.

[Gaede 98] Volker Gaede & Oliver Günther. *Multidimensional Access Methods*. ACM Comput. Surv., vol. 30, no. 2, pages 170–231, 1998.

[Gama 10] Joao Gama. Knowledge discovery from data streams. Chapman & Hall/CRC, 1st edition, 2010.

[Ganguly 07] A. R. Ganguly, Y. Fang O. A. Omintaomu, S. Khan & B. L. Bhaduri. Knowledge discovery from sensor data for scientific applications. Springer, 2007.

[Guha 03] Sudipto Guha, Adam Meyerson, Nina Mishra, Rajeev Motwani & Liadan O'Callaghan. *Clustering Data Streams: Theory and Practice.* IEEE Transactions on Knowledge and Data Engineering, vol. 15, no. 3, pages 515–528, May/June 2003.

[Guttman 84] Antonin Guttman. *R-Trees: A Dynamic Index Structure for Spatial Searching*. In SIGMOD Conference, pages 47–57, 1984.

[Jiang 02] Changhao Jiang, Yiheng Li, Minglong Shao & Peng Jia. *Accelerating Clustering Methods through Fractal Based Analysis*. In Proceddings of the ACM SIGKDD Workshop on Fractals and Self-similarity in Data Mining: Issues and Approaches, pages 16–20, Edmonton, Canada, July 2002.

[Jiang 06] Nan Jiang & Le Gruenwald. *Research issues in data stream association rule mining*. SIGMOD Record, vol. 35, no. 1, pages 14–19, 2006.

[Kantardzic 04] Mehmed Kantardzic, Pedram Sadeghian & Chun Shen. *The time diversification monitoring of a stock portfolio: an approach based on the fractal dimension*. In Proceedings of the 2004 ACM Symposium on Applied Computing(SAC'04), pages 637–641. ACM Press, March 2004.

[Kawale 12] Jaya Kawale, Snigdhansu Chatterjee, Dominick Ormsby, Karsten Steinhaeuser, Stefan Liess & Vipin Kumar. *Testing the significance of spatio-temporal teleconnection patterns*. In The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012, pages 642–650, 2012.

[Kifer 04] Daniel Kifer, Shai Ben-David & Johannes Gehrke. *Detecting Change in Data Streams.* In Proceedings of the Thirtieth International Conference on Very Large

Data Bases (VLDB'04), pages 180–191, Toronto, Canada, August/September 2004. Morgan Kaufmann.

[Korn 01] Flip Korn, Bernd-Uwe Pagel & Christos Faloutsos. *On the 'Dimensionality Curse' and the 'Self-Similarity Blessing'*. IEEE Transactions on Knowledge and Data Engineering, vol. 13, no. 1, pages 96–111, January/February 2001.

[Kosina 15] Petr Kosina & João Gama. *Very fast decision rules for classification in data streams*. Data Min. Knowl. Discov., vol. 29, no. 1, pages 168–202, 2015.

[Mandelbrot 83] Benoit B. Mandelbrot. The fractal geometry of nature: updated and augmented. W. H. Freeman and Company, New York, USA, 1983.

[Meng 06] Yu Meng & Margaret H. Dunham. *Mining Developing Trends of Dynamic Spatio-temporal Data Streams*. JCP, vol. 1, no. 3, pages 43–50, 2006.

[Nunes 10] Santiago A. Nunes, Luciana A. S. Romani, Ana M. H. Avila, Caetano Traina Junior, Elaine P. M. Sousa & Agma J. M. Traina. *Análise baseada em fractais para identificação de mudanças de tendências em múltiplas séries climáticas*. In Proceedings of the XXV Brazilian Symposium on Databases - Short Paper Session, pages 65–72, Belo Horizonte, Brasil, October 2010.

[Nunes 11a] S. A. Nunes, L. A. S. Romani, A. M. H. Avila, E. P. M. Sousa & A. J. M. Traina. *Combinando análise fractal e mineração de séries temporais para identificação de extremos climáticos regionais.* In Anais do XXX Concurso de Trabalhos de Iniciação Científica (CTIC 2011) - XXXI Congresso da Sociedade Brasileira de Computação (CSBC 2011), pages 164–173, Natal, Brasil, Julho 2011.

[Nunes 11b] S. A. Nunes, L. A. S. Romani, A. M. H. Avila, C. Traina, E. P. M. Sousa & A. J. M. Traina. *Fractal-based Analysis to Identify Trend Changes in Multiple Climate Time Series*. Journal of Information and Data Management (JIDM), vol. 2, pages 51–57, 2011.

[Nunes 12] Santiago A. Nunes, José E. M. Colabardini, Priscila P. Coltri, Ana M. H. Avila, Luciana A. S. Romani, Caetano Traina Junior, Agma J. M. Traina & Elaine P. M. Sousa. *ClimFractal Analyser: um ambiente de análise de séries temporais climáticas baseado em workflows*. In SBBD Demos, 2012.

[Pagel 00] Bernd-Uwe Pagel, Flip Korn & Christos Faloutsos. *Deflating the Dimensionality Curse Using Multiple Fractal Dimensions*. In Proceedings of the 16th International Conference on Data Engineering (ICDE'00), pages 589–598, San Diego, USA, February/March 2000.

[Papadimitriou 04] Spiros Papadimitriou, Anthony Brockwell & Christos Faloutsos. *Adaptive, unsupervised stream mining.* VLDB Journal, vol. 13, no. 3, pages 222–239, September 2004.

[Robinson 81] John T. Robinson. *The K-D-B-Tree: A Search Structure For Large Multidimensional Dynamic Indexes*. In Y. Edmund Lien, editeur, Proceedings of the 1981 ACM SIGMOD International Conference on Management of Data, Ann Arbor, Michigan, April 29 - May 1, 1981, pages 10–18. ACM Press, 1981.

[Rodrigues 07] Pedro Pereira Rodrigues & João Gama. *A Semi-Fuzzy Approach for Online Divisive-Agglomerative Clustering*. In Proceedings of the Workshop on Knowledge Discovery from Ubiquitous Data Streams (in conjunction with 18th European Conference on Machine Learning (ECML'07) and the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'07)), pages 107–116, Warsaw, Poland, September 2007.

[Romani 09] L. A. S. Romani, E. P. M. Sousa, M. X. Ribeiro, J. Zullo, C. Traina & A. Traina. *Employing Fractal Dimension to Analyze Climate and Remote Sensing Data Streams*. In Proceedings of the First SIAM Workshop on Multimedia Data Mining (in conjunction with the SIAM Data Mining Conference - SDM 2009), pages 1–12, Sparks, USA, 2009.

[Samet 84] Hanan Samet. *The Quadtree and Related Hierarchical Data Structures*. ACM Comput. Surv., vol. 16, pages 187–260, June 1984.

[Schroeder 91] Manfred Schroeder. Fractals, chaos, power laws: Minutes from an infinite paradise. W. H. Freeman and Company, New York, USA, 1991.

[Sigrist 12] Fabio Sigrist, Hans R. Künsch & Werner A. Stahel. *A dynamic nonstationary spatio-temporal model for short term prediction of precipitation*. The Annals of Applied Statistics, vol. 6, no. 4, pages 1452–1477, 2012.

[Sousa 06a] E. P. M. Sousa. *Correlation identification using the fractal theory*. PhD thesis, University of São Paulo, 2006.

[Sousa 06b] Elaine P. M. Sousa, Marcela X. Ribeiro, Agma J. M. Traina & Caetano Traina. *Tracking the Intrinsic Dimension of Evolving Data Streams to Update Association Rules*. In Proceedings of the 3rd International Wokshop on Knowledge Discovery from Data Streams (in conjunction with the 23rd International Conference on Machine Learning (ICML'06)), Pittsburgh, USA, 2006.

[Sousa 07a] Elaine P. M. Sousa, Agma J. M. Traina, Caetano Traina & Christos Faloutsos. *Measuring evolving data streams' behavior through their intrinsic dimension*. New Gen. Comput., vol. 25, no. 1, pages 33–60, January 2007.

[Sousa 07b] Elaine Parros Machado Sousa, Caetano Traina, Agma Juci Machado Traina, Leejay Wu & Christos Faloutsos. *A Fast and Effective Method to Find Correlations among Attributes in Databases*. Data Mining And Knowledge Discovery, vol. 14, no. 3, pages 367 – 407, June 2007.

[Tan 11] Swee Chuan Tan, Kai Ming Ting & Fei Tony Liu. *Fast Anomaly Detection for Streaming Data*. In Toby Walsh, editeur, IJCAI, pages 1511–1516. IJCAI/AAAI, 2011.

[Traina Jr. 05] Caetano Traina Jr., Elaine Parros Machado Sousa & Agma Juci Machado Traina. Using fractals in data mining, volume 1, page 30. Wiley/IEEE Press, 02/2005 2005.

[Traina 00] Caetano Traina, Agma Traina, Leejay Wu & Christos Faloutsos. *Fast Feature Selection using Fractal Dimension*. In XV Brazilian Symposium on Databases (SBBD'00), pages 158–171, João Pessoa, Brazil, October 2000.

[Traina 02] Caetano Traina, Agma J. M. Traina, Christos Faloutsos & Bernhard Seeger. *Fast Indexing and Visualization of Metric Data Sets using* S*lim-Trees*. IEEE Transactions on Knowledge and Data Engineering, vol. 14, no. 2, pages 244–260, March/April 2002.