PLANO DE TRABALHO

Programa de Mestrado em Ciência da Computação
Universidade Federal de Uberlândia

Aluno: Rodolfo Martignon Sevilhano Mendes

Orientador: Prof. Dr. Humberto Luiz Razente

Título do Trabalho: Uma nova abordagem para aumento da escalabilidade na geração de agrupamentos hierárquicos de séries espaço-temporais

Título do Trabalho: Efeitos da Amostragem no Agrupamento Hierárquico de Séries Temporais

Data de Início como Aluno Regular: 11/08/2014

Previsão da Defesa: Fevereiro de 2017

1 Introdução

A análise de séries temporais de imagens de satélite tem se mostrado uma importante ferramenta no estudo dos impactos das mudanças climáticas globais. A partir destas séries temporais, é possível analisar o uso e a cobertura do solo em uma determinada região, e como este uso varia ao longo do tempo.

Uma das possíveis técnicas para extrair conhecimento destas imagens é o agrupamento de séries temporais. A partir das imagens de satélite, métricas como temperatura da superfície, índice de reflexão e índice de vegetação são extraídos para cada pixel da imagem, para cada instante no tempo. O agrupamento das séries temporais obtidas, combinado com a visualização em mapa a quais grupos cada série pertence, permite ao tomador de decisão entender como o uso do solo está distribuído em uma região.

Por sua vez, o avanço das tecnologias de sensoreamento têm permitido que satélites coletem imagens com resoluções cada vez maiores. Isso significa que os conjuntos de dados são representados por um número cada vez maior de séries temporais, na ordem de milhões e até mesmo bilhões de séries para serem agrupadas.

Por meio do algoritmo k-médias é possível agrupar dados com complexidade de tempo linear, $O(n)$. Porém, esse algoritmo possui algumas desvantagens com relação ao significado dos agrupamentos gerados. Uma destas desvantagens é a necessidade do usuário fornecer antecipadamente o número de agrupamentos desejados através do parâmetro k. Ou seja, é necessário que o usuário tenha algum conhecimento prévio sobre o domínio do problema e que ele tenha uma estimativa do número de grupos existente no conjunto

de dados. Outra limitação está relacionada ao formato dos agrupamentos no espaço dos atributos. No k-médias, cada instância do conjunto de dados é atribuída ao grupo cujo centro esteja mais próximo, que por sua vez é recalculado a cada iteração. Assim, os agrupamentos tendem a ter formatos esféricos ou regiões densas são divididas por hiperplanos separadores projetados, o que nem sempre corresponde à melhor descrição dos dados.

Uma abordagem alternativa são os algoritmos hierárquicos aglomerativos. Nestes algoritmos, cada instância do conjunto de dados é colocada em seu próprio grupo inicialmente. Em seguida, os agrupamentos são aglutinados em grupos maiores, formando uma hierarquia de grupos. Diferentemente do k-médias, o usuário não precisa fornecer o número k de grupos previamente. Embora a saída do algoritmo hierárquico seja um dendograma (uma hierarquia de grupos), este pode ser convertido em uma partição de k grupos por meio de uma poda na árvore resultante. Por sua vez, as partições resultantes deste processo não estão limitadas a formatos esféricos, sendo o algoritmo capaz de detectar agrupamento com formatos arbitrários. Apesar dessas vantagens, os algoritmos hierárquicos tem alto custo computacional, de ordem $O(n^2)$ (quadrática) até $O(n^3)$ (cúbica) para tempo e espaço, tornando sua aplicação impraticável para grandes conjuntos de dados como séries temporais de imagens de satélites de alta resolução.

Neste trabalho uma nova abordagem é proposta para que se possa aproveitar as vantagens dos algoritmos hierárquicos mesmo em grandes conjuntos de dados. Esta abordagem consiste em, inicialmente, reduzir o conjunto de dados, e em seguida, aplicar o agrupamento hierárquico neste conjunto de dados reduzido. Por fim, as instâncias restantes são aglomeradas ao grupo mais próximo, utilizando a mesma medida de distância entre grupos utilizada no agrupamento inicial. Espera-se que, com esta abordagem, seja possível aplicar o agrupamento hierárquico em tempo consideravelmente menor, mas que a perda de qualidade dos agrupamentos resultantes seja mínima.

Esta proposta de trabalho está dividida da seguinte forma: na seção 1.1 estão listados o objetivo geral da pesquisa e os objetivos mais específicos. Na seção 1.2 encontra-se a formulação da hipótese a ser averiguada durante o trabalho e na seção 1.3 destaca-se a contribuição do trabalho. Na Seção 2, são apresentados os principais conceitos teóricos necessários para o desenvolvimento e entendimento do trabalho. Na Seção 3 é apresentada a proposta de algoritmo hierárquico aglomerativo baseada em redução de dados. Na Seção 5 são apresentados os resultados preliminares obtidos ao aplicar a técnica proposta. Na Seção 4, descreve-se a metodologia que será utilizada para o prosseguimento do trabalho. Por fim, na Seção 6 é apresentado o cronograma para a execução do restante da pesquisa.

1.1 Objetivos e Desafios de Pesquisa

O objetivo geral deste trabalho é desenvolver uma abordagem escalável para o agrupamento hierárquico de séries espaço-temporais de imagens de satélite que permita reduzir significativamente a complexidade de tempo de execução e espaço do algoritmo, sem no entanto reduzir a qualidade dos agrupamentos produzidos. Especificamente, deseja-se obter:

1. Desenvolvimento de uma abordagem para o agrupamento hierárquico aglomerativo de séries espaço-temporais incluindo uma etapa de pré-processamento baseada em redução de dados;

2. Desenvolvimento de uma abordagem para a geração de amostras de séries espaço-temporais por meio da análise de auto-similaridade;

3. Avaliação experimental das novas abordagens em termos de tempo de execução, consumo de memória e qualidade dos agrupamentos gerados visando o agrupamento dos diversos tipos de vegetação para identificação de áreas de plantio de culturas como a cana-de-açúcar.

A etapa 1 encontra-se concluída e os resultados experimentais iniciais encontram-se descritos na Seção 5. A etapa 2 encontra-se em final de implementação.

1.2 Hipótese

A hipótese deste projeto de pesquisa consiste em: "A redução do conjunto de dados de entrada por meio de uma amostragem baseada em fractais permite executar o agrupamento hierárquico aglomerativo em tempo menor, mantendo a qualidade dos agrupamentos produzidos".

1.2.1 Amostragem baseada em fractais

A hipótese de uso da teoria dos fractais para a geração de amostras significativas baseadas na análise da auto-similaridade vem do fato de que técnicas de amostragem ingênuas (*naive*), como amostragens aleatórias e amostragens estratificadas, não são adequadas para conjuntos de dados reais com presença de dados ruidosos e de exceções, uma vez que o comportamento desses algoritmos é imprevisível. Como em muitos conjuntos de séries temporais não há atributos de classe disponíveis (necessários nas abordagens clássicas de amostragens estratificadas), pretende-se fazer uma amostragem baseada na densidade da estrutura *box count*. O objetivo é obter um conjunto mínimo de instâncias que represente as características do conjunto original.

1.3 Contribuição

A contribuição esperada é uma estratégia para o agrupamento hierárquico aglomerativo que produza respostas mais rápidas para o usuário.

2 Revisão da Literatura Correlata

Nesta seção apresentaremos os principais conceitos teóricos relacionados ao trabalho desenvolvido. Na seção 2.1 é apresentado o conceito de séries temporais, que representam o objeto da mineração de dados deste trabalho. Na seção 2.2 é apresentado o processo de descoberta de conhecimento em banco de dados e suas principais etapas. Na seção 2.3 são listadas algumas das principais técnicas para amostragem de dados. Na seção 2.4 é apresentada a tarefa de análise de agrupamentos. Na seção 2.5, a abordagem hierárquica para agrupamentos é discutida em mais detalhes. São apresentados os algoritmos aglomerativo e divisivo, e são discutidas as principais medidas de distância entre grupos. Na seção 2.6 são apresentadas técnicas e índices para medir a qualidade dos agrupamentos

gerados, permitindo comparar objetivamente os resultados das diferentes abordagens de agrupamentos. Finalmente, em 2.7 apresenta a análise de fractais e a propriedade de auto-similaridade.

2.1 Séries Temporais

Uma série temporal é uma sequencia ordenada de valores coletados ao longo do tempo. Elas são usadas para medir fenômenos que variam com o tempo, e sua análise permite a construção de modelos que explicam estes fenômenos [Morettin and de Castro Toloi 2006]. São exemplos de séries temporais: o volume de poupança em um banco ao longo dos meses, a temperatura média de uma cidade durante a semana, ou ainda, o número de acessos a um *website* durante o dia.

Uma série temporal pode ser representada por uma função $Z(t)$, cujo domínio representa o tempo, e t, um determinado instante no tempo. Dada essa representação, uma série temporal pode ser classificada em discreta ou contínua, de acordo com o tipo de conjunto que representa o tempo [Morettin and de Castro Toloi 2006]. Por exemplo: o faturamento diário de um supermercado é uma série temporal discreta, pois o tempo é representado por um conjunto de intervalos enumeráveis, neste caso, dias. Por outro lado, a velocidade de um carro de corrida durante uma prova é um exemplo de série temporal contínua, pois a duração da corrida é um intervalo contínuo de tempo. Ainda, uma série temporal discreta pode ser obtida a partir de uma série contínua, bastando tomar amostras dos valores da série contínua em intervalos regulares de tempo.

No entanto, há casos em que são necessárias mais de uma variável para representar o fenômeno medido pela série temporal. Por exemplo, o *candlestick* diário de uma ação no mercado de capitais é formado por quatro valores: preço de abertura, preço mínimo, preço máximo e preço de fechamento. Nestes casos, a série temporal é classificada como multivariada, e é representada por uma função vetorial $Z(t) = [Z_1(t), Z_2(t), ...Z_r(t)]$, onde r é o número de variáveis que descrevem o fenômeno. Caso (Z_t) seja um escalar, a série temporal é classificada como univariada [Morettin and de Castro Toloi 2006].

2.2 Descoberta de Conhecimento em Bancos de Dados

A Descoberta de Conhecimento em Bancos de Dados, também conhecida pela sigla KDD (*Knowledge Discovery in Databases*) é o processo pelo qual dados brutos, coletados a partir das mais variadas fontes, são processados e transformados em informações úteis. Por sua vez, estas informações permitem o aprimoramento da tomada de decisão e até mesmo ampliação do conhecimento científico sobre um determinado fenômeno [Tan et al. 2009].

O processo de KDD envolve desde a aquisição dos dados até a disponibilização do conhecimento para o usuário final. De acordo com [Tan et al. 2009], este processo pode ser descrito pelas seguintes etapas:

1. Pré-processamento
2. mineração de dados
3. Pós-processamento

O objetivo da etapa de pré-processamento é preparar os dados que alimentarão a etapa de mineração de dados. Nesta etapa, podem ser realizadas uma série de tarefas que visam aumentar a qualidade dos dados fornecidos à mineração de dados. Na tarefa de *limpeza dos dados* são tratados atributos sem valor definido e ruídos. A tarefa de *integração de dados* consiste em consolidar fontes de dados de diversos tipos (arquivos de texto, planilhas, *web-services*, arquivos XML, bancos de dados) em uma única fonte de dados consolidada, usualmente um *data-warehouse*. A *redução da dimensionalidade* consiste em diminuir o número de atributos que serão considerados na mineração de dados. Dentre as principais técnicas podemos citar PCA (*Principal Component Analysis*) e DWT (*Discrete Wavelets Transforms*). Por fim, a *redução da numerosidade* busca representar o conjunto de dados através de um número reduzido de instâncias [Han et al. 2011].

O reconhecimento de padrões é efetivamente realizado na etapa de mineração de dados. As tarefas desta etapa são categorizadas de acordo com o conhecimento que se deseja extrair da base de dados analisada. Na tarefa de mineração de itens frequentes, deseja-se extrair de um banco de transações quais itens ocorrem conjuntamente com maior frequência. Na tarefa de classificação, o objetivo é inferir um modelo a partir do qual seja possível prever à qual classe uma determinada instância de dados pertence. Por fim, na análise de agrupamentos deseja-se descobrir a existência de grupos (*clusters*) de dados. Assim, é preciso que se estabeleça uma *medida de similaridade* entre as instâncias do banco de dados, de forma que se maximize a similaridade entre instâncias do mesmo grupo e se minimize a similaridade entre instâncias de grupos diferentes.

Por fim, na etapa de pós-processamento avalia-se se os padrões descobertos de fato representam um *conhecimento* novo sobre os dados. Para cada tipo de padrão descoberto, pode-se estabelecer uma *medida objetiva* sobre a qualidade do padrão [Han et al. 2011]. No caso dos agrupamentos, por exemplo, a qualidade destes pode ser medida em termos de *coesão* e *separação* [Tan et al. 2009].

Neste trabalho, será enfatizada a tarefa de agrupamento de dados, com atenção especial aos algoritmos hierárquicos de agrupamento. Também será discutido como as técnicas de redução de numerosidade influenciam o tempo de execução dos algoritmos hierárquicos e a qualidade dos agrupamentos produzidos, como etapa de pré-processamento.

2.3 Técnicas de Amostragem de dados

O objetivo das técnicas de amostragem de dados é reduzir o número de instâncias submetidas aos algoritmos de mineração de dados. Entre os desafios da amostragem de dados estão o balanceamento das instâncias com relação à ocorrência de instâncias raras ou de exceções. Considere um conjunto T com cardinalidade $|T| = N$. Entre as técnicas propostas na literatura destacam-se [García et al. 2015]:

- Amostragem aleatória de tamanho s sem substituição: criada pela escolha de s instâncias de T ($s < N$), onde a probabilidade de um exemplo ser escolhido é de $1/N$, de modo que todas as instâncias têm a mesma chance de serem escolhidas;

- Amostragem aleatória de tamanho s com substituição: semelhante à anterior, exceto pelo fato que a cada vez que uma instância é escolhida, permanece no conjunto e pode ser escolhida novamente;

- Amostragem balanceada: criada levando-se em consideração um conjunto de critérios pré-definidos, por exemplo, para manter a proporcionalidade de instâncias entre classes conhecidas;

- Amostragem de agrupamentos: escolha de grupos específicos resultantes de técnicas de agrupamento;

- Amostragem estratificada: obtida por meio da divisão de um conjunto T em partes mutualmente disjunta seguida da escolha de uma amostragem aleatória em cada divisão.

2.4 Análise de Agrupamentos

A análise de agrupamentos é uma tarefa de mineração de dados cujo objetivo é, automaticamente, particionar o conjunto de dados em subconjuntos chamados grupos. Os objetos reunidos em um mesmo grupo devem ser similares entre si, enquanto que objetos de grupos separados devem ser diferentes. Ao conjunto dos grupos resultantes da análise dá-se o nome de *agrupamento*.

A análise de agrupamentos pode ser usada como uma ferramenta para extração de conhecimento sobre um conjunto de dados ou então, como um etapa de pré-processamento para outras tarefas de mineração de dados. Por exemplo, em [Gonçalves et al. 2014], a análise de agrupamentos foi utilizada para identificar o uso do terreno em diferentes regiões do estado de São Paulo, Brasil. Já em [Petitjean et al. 2014], a análise de agrupamentos foi utilizada para eleger protótipos que posteriormente seriam utilizados como dados de treinamento para a tarefa de classificação 1-NN.

Existem diversas abordagens para o agrupamento de dados. No agrupamento por *particionamento* o conjunto de dados é dividido em k grupos, com cada grupo contendo pelo menos um objeto do conjunto. De maneira geral, estes algoritmos consistem em: a partir de um agrupamento inicial, iterativamente realocar os objetos em grupos mais significativos até que um critério de parada seja atingido. Podemos incluir nesta categoria os algoritmos *k-médias* e *k-medoids*.

Uma abordagem alternativa é o agrupamento *hierárquico*. Nesta abordagem, os objetos são organizados em uma hierarquia de grupos. Por sua vez, esta hierarquia pode ser construída por duas maneiras diferentes: *aglomerativa* e *divisiva*.

Na abordagem aglomerativa cada objeto de dados é inicialmente incluído em seu próprio grupo. Em seguida, cada grupo é aglomerado com o seu grupo mais próximo, formando uma relação "pai-filho" entre o grupo resultante e os grupos menores. Esse processo se repete até que um único grupo, que contenha todos os dados do conjunto, seja obtido. Já na abordagem divisiva o processo se inverte. Todos os objetos de dados são agrupados em um único grupo inicial, que será a raiz da hierarquia. Por sua vez, este grupo inicial é sucessivamente dividido em grupos menores, até que cada objeto esteja em seu próprio grupo.

Na seção 2.5 a abordagem hierárquica será explorada com mais detalhes.

2.5 Abordagem Hierárquica para Agrupamentos

Na abordagem hierárquica de agrupamento, os grupos são organizados em árvore, de forma que cada nodo desta árvore representa um grupo. Desta maneira, estabelece-se uma relação pai-filho entre os grupos, tal que, dado um grupo pai C_p que tenha n filhos $\{C_1, C_2, ..., C_n\}$, então $C_i \subset C_p$ para todo $1 \leq i \leq n$, como mostra a Figura 1(b). Nas folhas desta árvore, encontram-se as instâncias de dados, cada uma incluída em seu próprio grupo. Por sua vez, na raiz encontra-se o grupo que abrange todo o conjunto de dados. A Figura 1(a) mostra a hierarquia entre os grupos através de um diagrama conhecido como *dendrograma*.

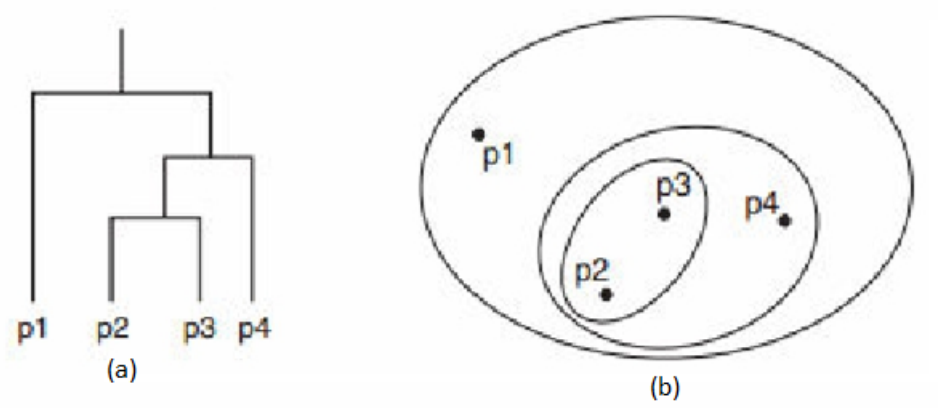

Figura 1: Exemplo de agrupamento hierárquico [Tan et al. 2009].

A organização de grupos em hierarquias tem aplicações em diversas áreas. Em recuperação da informação por exemplo, documentos podem ser agrupados de maneira hierárquica de acordo com o assunto, revelando uma estrutura de tópicos e sub-tópicos. Em Biologia, o agrupamento de espécies de acordo com suas características pode ajudar na compreensão sobre como essas espécies evoluíram ao longo do tempo.

Uma das principais vantagens da abordagem hierárquica de agrupamento é que não é necessário que o usuário informe o número k de grupos desejado previamente. Porém, nem sempre é interessante para o usuário analisar todos os grupos obtidos pelo agrupamento hierárquico. Nestes casos, a hierarquia original pode ser convertida em um particionamento através do procedimento de poda (*prunning*).

Nesta seção serão discutidos os principais aspectos relativos ao agrupamento hierárquicos. Serão abordados: algoritmo AGNES para agrupamento aglomerativo, algoritmo DIANA para agrupamento hierárquico divisivo e as principais medidas de distância entre clusters: single linkage, complete linkage, average linkage, mean distance e Ward's distance. Por fim serão analisados vantagens e desvantagens da abordagem hierárquica.

2.5.1 AGNES: algoritmo aglomerativo

O algoritmo AGNES (*AGglomerative NESting*) é o algoritmo elementar para executar o agrupamento hierárquico aglomerativo. Basicamente, seu procedimento consiste alocar inicialmente cada instância de dados em seu próprio grupo e então, sucessivamente, fundir

os grupos mais próximos entre si, até que todos os objetos sejam aglomerados em um único grupo, como mostra o Algoritmo 1.

Algoritmo 1: AGNES

Entrada: conjunto de dados X
Saída: agrupamento hierárquico
$n \leftarrow |X|$
$\mathcal{C} \leftarrow \{C_i = \{x_i\} \,|x_i \in X\}$
$M \leftarrow (m_{ij})_{n \times n} \mid m_{ij} = dist\,(C_i, C_j)\,, C_i, C_j \in \mathcal{C}$
repita
  encontra o par de grupos mais próximos C_i e C_j
  $C \leftarrow C_i \cup C_j$
  $\mathcal{C} \leftarrow (\mathcal{C} \setminus \{C_i, C_j\}) \cup \{C\}$
  recalcula M
até $C \equiv X$

Após cada instância de dados ser atribuída ao seu próprio grupo, a matriz de distâncias M é calculada, armazenando a distância de cada par de grupos existente. Então, sucessivamente, o algoritmo localiza a menor entrada na matriz de distância M, que equivale a encontrar o par dos grupos mais próximos, aglomera os grupos encontrados e recalcula a matriz de distância M.

Dado o algoritmo básico do agrupamento hierárquico aglomerativo, a principal diferença entre as diferentes abordagens nesta categoria são as medidas de similaridade entre grupos usadas. Por sua vez, estas podem ser baseadas em grafos ou baseadas em protótipos [Tan et al. 2009]. São medidas de distância baseadas em grafos:

- **Ligação simples**: A ligação simples, ou *single linkage*, toma como similaridade entre dois grupos a distância entre seus elementos mais próximos, e é calculada pela Equação 1:

$$dist\,(C_i, C_j) = \min_{x_i \in C_i, x_j \in C_j} \{|x_i - x_j|\} \tag{1}$$

  Ao tomar-se as instâncias de dados como vértices de um grafo, e as ligações entre grupos como vértices ponderados, então o agrupamento gerado é correspondente a uma *árvore geradora mínima* [Han et al. 2011], de forma que os grupos formados tendem a ser contíguos no espaço dos atributos [Tan et al. 2009].

- **Ligação completa**: A ligação completa, ou *complete linkage* é a medida oposta à ligação simples, pois toma como similaridade entre dois grupos a distância entre seus elementos mais distantes, sendo calculada pela Equação 2:

$$dist\,(C_i, C_j) = \max_{x_i \in C_i, x_j \in C_j} \{|x_i - x_j|\} \tag{2}$$

- **Distância Média**: Por fim, a similaridade entre dois grupos pode ser medida através da distância média entre os pares dos grupos (*group average*). Essa medida é um balanceamento entre a ligação simples e a ligação completa, e é obtida pela média das distâncias entre cada um dos pares ordenados (x_i, x_i), com $x_i \in C_i$ e $x_i \in C_j$. É obtida pela Equação 3:

$$dist\,(C_i, C_j) = \frac{1}{n_i n_j} \sum_{x_i \in C_i} \sum_{x_j \in C_j} |x_i - x_j| \tag{3}$$

Além das medidas baseadas em grafos, também podem ser aplicadas medidas baseadas em protótipos. São elas:

- **Distância entre centroides**: a distância entre centroides toma como medida de similaridade entre grupos a distância entre as médias dos grupos. Seja o centroide μ_i de um grupo o seu objeto médio, dados pela Equação 4:

$$\mu_i = \frac{1}{|C_i|} \sum_{x_i \in C_i} x_i \tag{4}$$

Então, a distância entre os grupos é dada pela distância entre os respectivos centroides, como mostra a Equação 5:

$$dist\,(C_i, C_j) = |\mu_i - \mu_j| \tag{5}$$

- **Método de Ward**: o método de Ward também considera o centroide como representante do grupo todo. Porém, a distância entre grupos é medida pela variação da soma dos erros quadrados, ou SSE (*squared sum of errors*), que se obtém ao aglomerar dois grupos C_i e C_j. Seja a soma dos erros quadrados de um grupo i, dada pela Equação 6:

$$SSE_i = \sum_{x \in C_i} |x - \mu_i|^2 \tag{6}$$

Então, a distância de Ward será dada pela variação da soma dos erros quadrados obtida ao aglomerar os grupos C_i e C_j:

$$dist(C_i, C_j) = \Delta SSE_{ij} = SSE_{ij} - SSE_i - SSE_j \tag{7}$$

Em [Zaki and Meira 2014], é demonstrado que o método de Ward corresponde a uma versão ponderada da distância entre centroides:

$$dist(C_i, C_j) = \Delta SSE_{ij} = \left(\frac{n_i n_j}{n_i + n_j}\right) |\mu_i - \mu_j|^2 \tag{8}$$

A primeira etapa do algoritmo AGNES envolve o cálculo da matriz de distâncias entre os grupos iniciais. Dado o número n de instâncias no conjunto de dados, o cálculo da distância entre todos os pares de objetos envolve $\frac{n(n-1)}{2}$ cálculos, considerando $dist(C_i, C_j) = dist(C_j, C_i)$. Assim, o cálculo da matriz de distâncias tem complexidade de tempo e espaço $O(n^2)$.

Na etapa de aglomeração, o laço principal do algoritmo deve ser executado $n-1$ vezes. A cada iteração i, a busca pelos pares mais próximos leva tempo proporcional a $(n-i)^2 = O(n^2)$, assim como o recálculo da matriz de distâncias. Portanto, o algoritmo AGNES tem complexidade de tempo $O(n^3)$ [Tan et al. 2009]. No entanto, com o emprego de estruturas de dados como o *heap* é possível buscar a menor distância em tempo $O(1)$, enquanto que a atualização do *heap* com as novas distâncias calculadas leva $O(\log n)$. Portanto, a complexidade de tempo do algoritmo nessas condições é $O(n^2 \log n)$ [Zaki and Meira 2014].

2.5.2 DIANA: algoritmo divisivo

O algoritmo DIANA (DIvisive ANAlysis) é uma abordagem divisiva para o agrupamento hierárquico [Kaufman and Rousseeuw 1990]. Ao contrário da abordagem aglomerativa, na abordagem divisiva o conjunto de dados é tomado como o grupo inicial do agrupamento. Então, este grupo inicial é sucessivamente dividido em grupos menores, até que cada instância de dados esteja em seu próprio grupo. O algoritmo 2 mostra este procedimento.

Algoritmo 2: DIANA

Entrada: conjunto de dados D
Saída: agrupamento hierárquico
$\mathcal{C} \leftarrow \{D\}$
repita
  para cada $C \in \mathcal{C} \mid |C| > 1$ **faça**
    $\mathcal{C} \leftarrow \mathcal{C} \cup Divide(C)$
    $\mathcal{C} \leftarrow \mathcal{C} \setminus \{C\}$
  fim
até $|C| = 1, \forall C \in \mathcal{C}$

Devido aos desafios que a abordagem divisiva impõe, esta é pouco estudada na literatura. Um dos principais problemas encontrados na abordagem divisiva é justamente a divisão do grupo inicial. Na abordagem aglomerativa, todas as possíveis fusões são consideradas no passo inicial, dado que, para um conjunto de dados com n objetos, existem C_n^2 pares possíveis, como mostra a equação 9:

$$C_n^2 = \frac{n(n-1)}{2} \tag{9}$$

Porém, para a abordagem divisiva enumerar todas as possíveis divisões de n objetos em dois grupos não vazios, deverão ser consideradas $2^{n-1} - 1$ possibilidades, tornando a análise impraticável até mesmo para pequenos valores de n.

Dada a impossibilidade de se analisar todas as possíveis divisões do grupo inicial, o algoritmo DIANA utiliza uma heurística simples para dividir os grupos. Dado um grupo

C_i com n objetos, calcula-se a dissimilaridade de cada objeto $x \in C_i$ através da distância média do objeto x aos objetos restantes, como mostra a Equação 10

$$\mu_{xi} = \frac{1}{n-1} \sum_{y \in C_i} |x - y| \tag{10}$$

Para cada grupo que será dividido, mede-se a dissimilaridade de cada objeto com relação ao resto do grupo. Então, o mais dissimilar é encontrado e colocado em um grupo separado (*splinter group*). Novamente, as dissimilaridades são calculadas. Porém, também são calculadas as dissimilaridades entre os objetos restantes e os objetos do grupo separado. Caso algum objeto esteja mais próximo do grupo separado, este é transferido. Esse processo continua até que não haja mais transferências, e ao seu final, obtém-se a divisão em dois grupos. A Função Divide apresenta este procedimento.

Função Divide(C_{ij}: grupo a ser dividido)

```
C_j ← {x ∈ C_ij | μ_xij > μ_yij, ∀y ≠ x}
C_i ← C_ij \ C_j
transferiu ← true
enquanto transferiu = true faça
    transferiu ← false
    para cada x ∈ C_i faça
        recalcula μ_xi
        recalcula μ_xj
    fim
    x ← x ∈ C_i | μ_xi > μ_yi, ∀y ≠ x
    se μ_xi > μ_xj então
        C_j ← C_j ∪ {x}
        C_i ← C_i \ {x}
        transferiu ← true
    fim
fim
retorna {C_i, C_j}
```

2.6 Validação de Agrupamentos

A validação de agrupamentos é uma tarefa de pós-processamento da descoberta de conhecimento em bancos de dados. Dadas as diferentes abordagens para se produzir agrupamentos, e a variabilidade dos parâmetros necessários a esses algoritmos, surge a necessidade de se estabelecer critérios objetivos para avaliar e comparar os resultados dos algoritmos de agrupamento. Existem diversas métricas propostas na literatura para validar a qualidade de agrupamentos. Elas são dividas basicamente em dois grupos:

- **Externas**: o agrupamento é avaliado segundo critérios externos ao agrupamento. Um exemplo seria o conhecimento prévio sobre a categoria de cada uma das instâncias do conjunto de dados;

- **Internas**: neste caso, o agrupamento é avaliado a partir de características inerentes ao próprio agrupamento resultante. Duas medidas bastante comuns são a coesão e a separação dos agrupamentos.

Nas próximas sessões, serão apresentadas as principais métricas que poderão ser utilizadas neste trabalho de pesquisa.

2.6.1 Pureza

A pureza é uma medição externa da qualidade de um agrupamento. Para medi-la, é necessário que a cada objeto do conjunto de dados esteja associada uma categoria. A pureza mede a uniformidade dos grupos obtidos. Quanto menor o número de categorias presentes uma cada grupo, maior será a pureza do agrupamento. Para um determinado grupo C_i, a sua pureza é calculada pela Equação 11:

$$pureza_i = \frac{1}{n_i} \max_{j=1}^{k} \{n_{ij}\}, \tag{11}$$

onde n_i é o número de objetos no grupo C_i e n_{ij} é o número de objetos pertencentes à categoria j que estão presentes no grupo C_i. Por sua vez, a pureza do agrupamento é obtida pela média ponderada das purezas de cada um dos r grupos, como mostra a Equação 12.

$$pureza = \sum_{i=1}^{r} \frac{n_i}{n} pureza_i \tag{12}$$

A pureza de um agrupamento pode assumir valores entre 0 e 1, sendo que o valor 1 indica o grau máximo de pureza para um agrupamento. Destaca-se o fato de que é possível um agrupamento obter $pureza = 1$ mesmo quando o número de grupos é maior que o número de categorias. Nesse caso, os grupos são subconjuntos de cada categoria.

2.6.2 Índice Dunn

O índice Dunn é uma medida interna da qualidade de um agrupamento. Ele mede a qualidade do agrupamento como a razão entre a separação entre os grupos e a coesão dos objetos de um mesmo grupo. Seu valor pode variar entre 0 e $+\infty$, de forma que quanto maior o valor do índice Dunn, melhor será a qualidade do agrupamento.

Este índice é baseado em uma visão de grafos dos agrupamentos. A separação entre os grupos é representada pela menor distância entre objetos de grupos diferentes, enquanto que a coesão é medida pelo maior diâmetro de um grupo, como é mostrado pelas respectivas distâncias A e B na Figura 2.

2.6.3 Índice Davies-Boulding

O índice Davies-Boulding é também uma medida interna da qualidade de um agrupamento. Assim como o índice Dunn, o índice Davies-Boulding reúne em uma grandeza escalar tanto a coesão quanto a separação de um agrupamento. No entanto, este índice é baseado em uma representação de protótipos dos agrupamentos.

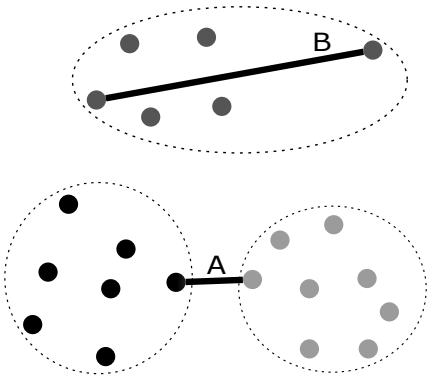

Figura 2: Separação e coesão de grupos no cálculo do índice Dunn.

Seja μ_i o centroide de um grupo, denotado pela Equação 4. O desvio-padrão desse grupo é obtido pela Equação 13.

$$\sigma_i = \sqrt{\frac{\sum_{x \in C_i} \left(|x - \mu_i|\right)^2}{n_i}} \tag{13}$$

Dado um par de grupos C_i e C_j, o valor do índice Davies-Bouldin para estes grupos é dado pela Equação 14.

$$DB_{ij} = \frac{\sigma_i + \sigma_j}{|\mu_i - \mu_j|} \tag{14}$$

Por fim, o valor do índice Davies-Boulding para o agrupamento todo é o valor máximo do índice entre os pares normalizado pelo número k de grupos, como mostra a Equação 15.

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} \{DB_{ij}\} \tag{15}$$

Assim, dada a representação baseada em protótipos do índice Davies-Boulding, observa-se que a separação entre dois grupos é dada pela distância entre seus centróides, enquanto que a coesão dos grupos é representada pela soma de seus respectivos desvios-padrão. Assim, quanto menor o valor do índice Davies-Boulding para um par de grupos, melhor será a qualidade do agrupamento.

2.7 Fractais e a propriedade de auto-similaridade

Um fractal pode ser definido pelo conceito de auto similaridade, no qual partes de qualquer tamanho de um fractal são similares (exata ou estatisticamente) ao conjunto todo. Um exemplo clássico de um fractal criado por meio da construção repetitiva é o triângulo de Sierpinski, construído por meio de um processo iterativo, onde se retira de um triângulo o triângulo central e para cada triângulo resultante realiza-se o mesmo processo, recursivamente, conforme apresentado na Figura 3. O triângulo de Sierpinski apresenta características interessantes, como o fato de cada triângulo interior ser uma miniatura do

triângulo em que está inserido, perímetro tendendo ao infinito e área tendendo a zero quando o número de iterações tende ao infinito [Schroeder 1991].

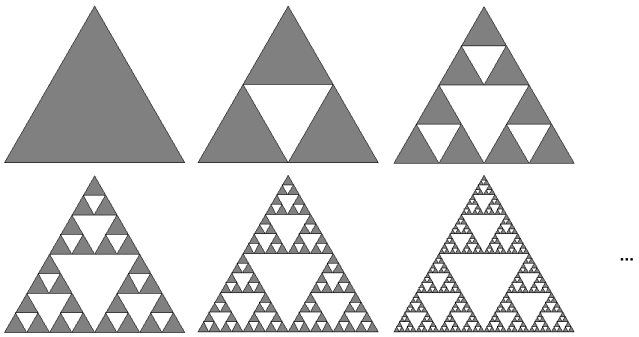

Figura 3: Construção do triângulo de Sierpinski [Schroeder 1991].

O conceito de auto similaridade está relacionado com periodicidade. Muitos fenômenos naturais e humanos acontecem com periodicidade. São exemplos o uso do solo pela agricultura, o valor das moedas e das ações, o comportamento de redes de comunicação, as filas de caixa de supermercado, o uso das estradas, as músicas que tocam nas rádios e os efeitos da economia na vida das pessoas. Uma razão para a universalidade dessas movimentações harmônicas é a linearidade aproximada de muitos sistemas e a sua invariância com deslocamento no espaço e tempo. Experimentos realizados com alguns conjuntos de dados sintéticos e reais mostram que os dados referentes aos fenômenos humanos caracterizam-se por apresentarem uma distribuição fractal [Traina-Jr. et al. 2010].

Para a análise da auto similaridade de conjuntos contendo fenômenos naturais e humanos, chamados de fractais estatisticamente auto-similares, utiliza-se o método *Box-Counting*. Para encontrar o *Box-Counting* de um conjunto de dados imerso em um espaço E-dimensional, deve-se dividir esse espaço em células de um hipercubo de lado r, recursivamente, até encontrar um elemento por célula [Traina-Jr. et al. 2010]. O método foi proposto para cálculo da dimensão fractal, que corresponde ao número mínimo de dimensões para representação de um conjunto (dimensionalidade intrínseca).

3 Agrupamento hierárquico aglomerativo de séries espaço-temporais

Nesta seção será descrita a técnica proposta para agrupamento hierárquico de séries-temporais. Basicamente, esta técnica consiste em reduzir o tamanho do conjunto de dados por meio de técnicas de amostragem, aplicar o agrupamento aglomerativo, e finalmente, atribuir as instâncias restantes aos seus grupos mais próximos, como mostra o Algoritmo 3.

A redução de dados tem papel fundamental na obtenção da escalabilidade dos algoritmos de agrupamento. Assim, o verdadeiro desafio na aplicação das técnicas de redução de dados é manter as características do conjunto de dados original, para que seja possível descobrir a estrutura dos grupos presentes.

Algoritmo 3: Agrupamento Hierárquico Aglomerativo com Amostragem

Entrada: Conjunto de dados D, número de grupos K
Saída: Agrupamento $\mathcal{C}$
1. Selecionar amostra $\mathcal{D}$, tal que $|\mathcal{D}| < |D|$
2. Obter o agrupamento hierárquico $\mathcal{H} \leftarrow AGNES(\mathcal{D})$
3. Aplicar o procedimento de poda em $\mathcal{H}$, obtendo o agrupamento $\mathcal{C}_0$, com K grupos
4. Atribuir os objetos restantes em $D \setminus \mathcal{D}$ aos grupos em $\mathcal{C}_0$, obtendo o agrupamento final $\mathcal{C}$

A redução do tamanho do conjunto de dados por meio de amostragem aleatória é uma técnica utilizada em vários algoritmos de agrupamento, entre eles CURE [Guha et al. 1998], CLARA [Kaufman and Rousseeuw 1990] e YADING [Ding et al. 2015]. Nestes trabalhos, a amostragem aleatória é aplicada estimando-se o tamanho mínimo das amostras para que as características dos grupos sejam preservadas.

3.1 Amostragem Baseada em Fractais

Em vários trabalhos da literatura científica da área a análise de fractais mostrou-se uma técnica promissora na redução da dimensionalidade dos dados. No entanto, até o momento, não foram encontrados trabalhos que tenham aplicado técnicas de análise de fractais à redução de conjuntos de dados espaço-temporais.

Neste trabalho propõe-se a aplicação da técnica de *Box-plot counting* para efetuar a redução de dados. Esta técnica permite detectar a auto-similaridade dos dados pela contagem das células que contêm um ou mais pontos. Essa característica será usada para a detecção de ruídos e *outliers*, permitindo a escolha de amostras estratificadas pela densidade dos hiper-retângulos em vários níveis de resolução.

4 Métodos de Pesquisa

O algoritmo proposto na seção 3 foi implementado em linguagem R[1] e C/C++, de modo que podem ser usadas diferentes técnicas de amostragem e diferentes medidas de distância, entre elas a distância Euclidiana e a *Dynamic Time Warping*. O algoritmo proposto vem sendo avaliado em termos do balanceamento do tempo de execução e da qualidade dos agrupamentos gerados. O algoritmo será avaliado com diferentes tamanhos de amostras, número de grupos e medidas de distância. Para cada número de grupos e medida de distância, a versão do agrupamento hierárquico com amostragem será comparada ao agrupamento hierárquico convencional, isto é, com todos os dados do conjunto e também com o agrupamento gerado aleatoriamente.

[1] https://www.r-project.org

4.1 Conjuntos de Dados

Os conjuntos de dados utilizados nos experimentos foram processados e fornecidos pela EMBRAPA [emb 2016] e têm resolução de 1 km/pixel. Na próxima etapa dos experimentos serão utilizados conjuntos de dados provenientes de imagens com resolução de 250 m/pixel. Outros conjuntos disponíveis no repositório em [glo 2016] também serão avaliados.

5 Resultados Preliminares

Nesta seção serão descritos os experimentos preliminares cujo intuito foi comparar o agrupamento hierárquico por amostragem, proposto na seção 3 e o algoritmo AGNES, que constrói o agrupamento hierárquico aglomerativo com todo o conjunto de dados.

O conjunto de dados utilizado nos testes são valores de índice NDVI (*Normalized Difference Vegetation Index*) de uma área localizada entre as latitudes $-8,55$ e $-8,45$, e as longitudes $-38,25$ e $37,25$, que corresponde a uma região do estado de Pernambuco, Brasil. Estes dados foram extraídos a partir de imagens coletadas por um satélite MODIS (*Moderate Resolution Imaging Spetroradiometer*) durante o ano de 2003, em períodos de 16 dias. Assim, a cada *pixel* da imagem é associada uma série temporal composta pelos valores do índice NDVI coletados ao longo do ano. O conjunto de dados utilizado possui um total de 9812 séries temporais, cada uma 23 coletas do índice NDVI. A tabela 1 mostra um subconjunto destes dados.

Latitude	**Longitude**	**01/01/2003**	**17/01/2003**	**02/02/2003**
-8.36	-40.89	0.76	0.83	0.76
-8.36	-40.89	0.71	0.83	0.76
-8.36	-40.89	0.73	0.81	0.62
-8.36	-40.89	0.68	0.74	0.72
-8.36	-40.89	0.69	0.74	0.72

Tabela 1: Amostras do conjunto de dados MODIS-PE-2003.

O experimento consistiu em comparar o algoritmo hierárquico aglomerativo por amostragem com o algoritmo AGNES, ambos utilizando a ligação simples (*single linkage*) como medida de distância entre grupos. Os algoritmos foram comparados em relação ao seu tempo de execução e à qualidade dos agrupamentos gerados. Para medir a qualidade dos agrupamentos, utilizou-se o índice Dunn e o índice Davies-Bouldin.

Para cada um dos algoritmos comparados, foi realizada a poda da hierarquia gerada em grupos $k = 3, 5, 7$. Por ser determinístico, o algoritmo AGNES foi executado uma única vez para cada valor de k. Já o algoritmo por amostragem foi executado com amostras de tamanho $m = 10, 100, 1000$. Por sua natureza probabilística, o algoritmo por amostragem foi executado 10 vezes para cada combinação de m e k, e foram calculadas as médias do tempo de execução e das métricas de qualidade dos agrupamentos.

Na tabela 2, são mostrados as médias do tempo de execução obtidos para cada tamanho de amostra (a amostra de tamanho 9812 corresponde à execução do algoritmo AGNES).

Tamanho da Amostra	Tempo Médio de Execução (segundos)
10	0.03
100	0.06
1000	2.68
9812	3416.65

Tabela 2: Tempo de execução por tamanho de amostra.

Embora o algoritmo baseado em amostragem precise, ao final do algoritmo, atribuir as instâncias restantes ao grupo mais próximo, o tempo de execução dessa etapa é da ordem de $O(nm)$, ou seja, linear em relação ao tamanho n do conjunto de dados. Assim, diminuir o número de instâncias processadas pelo algoritmo aglomerativo representou uma diminuição significativa no tempo de execução do agrupamento, como esperado. O gráfico da Figura 4 mostra a comparação dos tempos de execução em escala logarítmica.

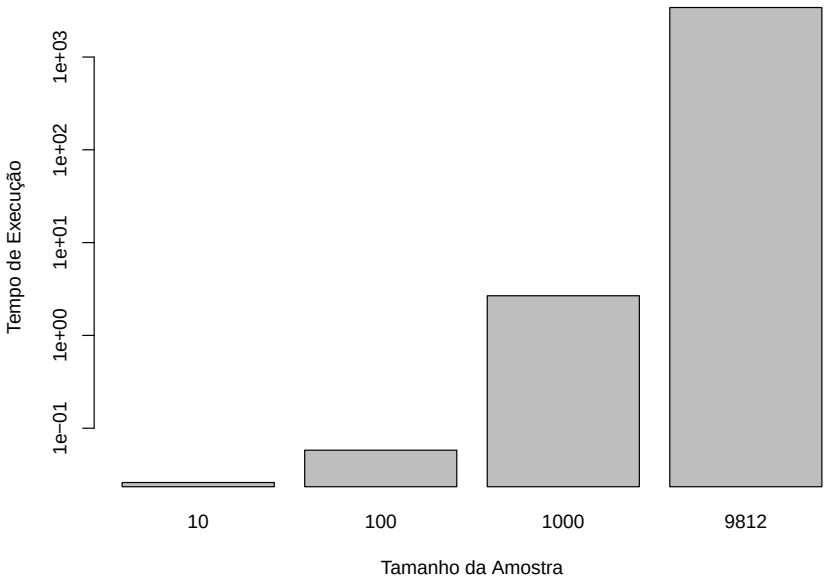

Figura 4: Tempo de execução por tamanho de amostra

Para comparar a qualidade dos agrupamentos produzidos foram utilizados os índices Dunn e Davies-Boulding. Além do tamanho da amostragem, também foi avaliado se o número de grupos utilizados na poda influenciaria a qualidade dos agrupamentos produzidos. Os resultados obtidos para o índice Dunn são apresentados na Tabela 3.

O índice Dunn é obtido pela razão entre a menor distância entre pares de grupos diferentes e a maior distância entre pares de um mesmo grupo. Assim, para agrupamentos de maior qualidade, o valor desse índice é maior. Como pode ser observado no gráfico da Figura 5, o número de grupos utilizado na poda influenciou pouco a qualidade dos agrupamentos obtidos. Por sua vez, o fator determinante na qualidade foi o tamanho da amostra utilizada nos algoritmos.

Tamanho da Amostra	Grupos	Índice Dunn
10	3	0.03
	5	0.02
	7	0.02
100	3	0.03
	5	0.03
	7	0.02
1000	3	0.04
	5	0.03
	7	0.03
9812	3	0.21
	5	0.20
	7	0.19

Tabela 3: Índice Dunn por tamanho de amostra e número de grupos.

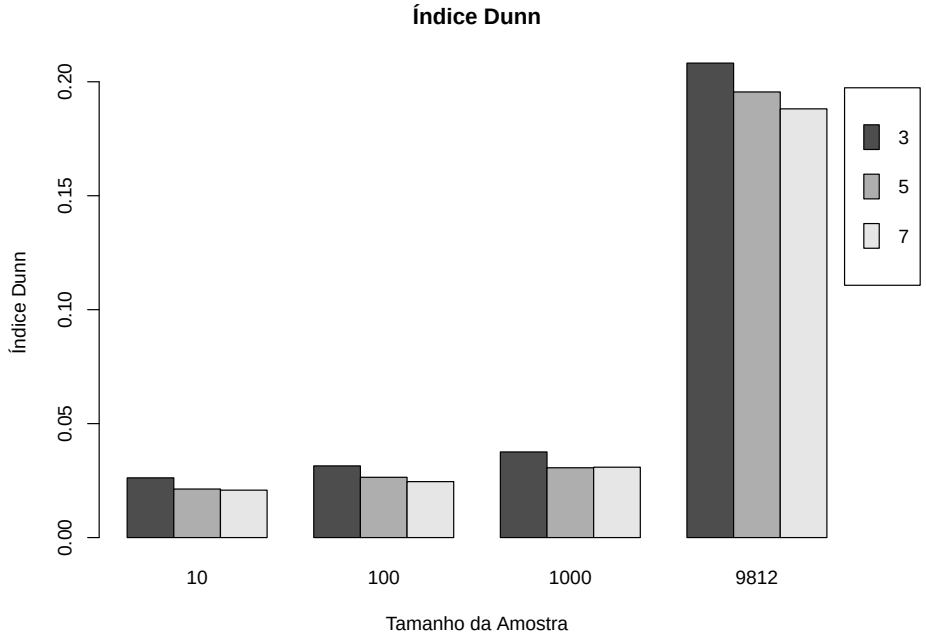

Figura 5: Índice Dunn por tamanho da amostra e número de grupos

Houve grande diferença entre a qualidade dos agrupamentos produzidos pelo algoritmo AGNES e os agrupamentos produzidos pelo algoritmo baseado em amostragem. Dado que o índice Dunn é influenciado pela menor distância entre pares de grupos diferentes, esse desempenho pode ser explicado pelo fato do algoritmo AGNES garantir que os pares de objetos mais próximos serão colocados no mesmo grupo. Por outro lado, no algoritmo baseado em amostragem, existe a possibilidade de que objetos muito próximos sejam colocados em grupos separados, afetando negativamente seu desempenho.

A qualidade inferior dos agrupamentos produzidos pelo algoritmo baseado em amostragem também foi refletida no índice Davies-Boulding. Este índice é obtido para cada par de grupos como a relação entre a soma dos desvios-padrão e a distância entre as médias dos grupos. Assim, quanto maior a qualidade dos agrupamentos, menor será o valor desse índice, pois menor será a dispersão dentro de um grupo e maior será a distância entre seus centros.

Tamanho da Amostra	**Grupos**	**Índice Davies-Bouldin**
10	3	67.11
	5	75.50
	7	76.94
100	3	49.32
	5	50.61
	7	54.27
1000	3	34.51
	5	37.51
	7	36.51
9812	3	4.77
	5	4.93
	7	5.30

Tabela 4: Índice Davies-Bouldin por tamanho de amostra e número de grupos.

Os dados mostrados na Tabela 4 também indicam que a qualidade dos agrupamentos produzidos pelo algoritmo baseado em amostragem foi inferior. Isso também pode ser explicado pela garantia que o algoritmo AGNES oferece de agrupar os pares mais próximos já no primeiro passo do algoritmo, pois diminui a dispersão dos dados. O gráfico da Figura 6 mostra que, com relação ao índice Dunn, este índice sofreu maior variação de acordo com o número de grupos e o tamanho das amostras.

5.1 Conclusões Preliminares

Por meio destes experimentos confirmou-se que o algoritmo proposto na Seção 3 permite que o agrupamento hierárquico aglomerativo seja executado em tempo muito menor que o algoritmo original. Porém, a amostragem aleatória impactou negativamente na qualidade dos agrupamentos obtidos.

No entanto, ainda não é possível descartar a utilização das técnicas de redução de dados para obter melhor desempenho do agrupamento hierárquico aglomerativo. Ao longo do trabalho proposto, pretende-se ampliar o escopo dos experimentos, averiguando o

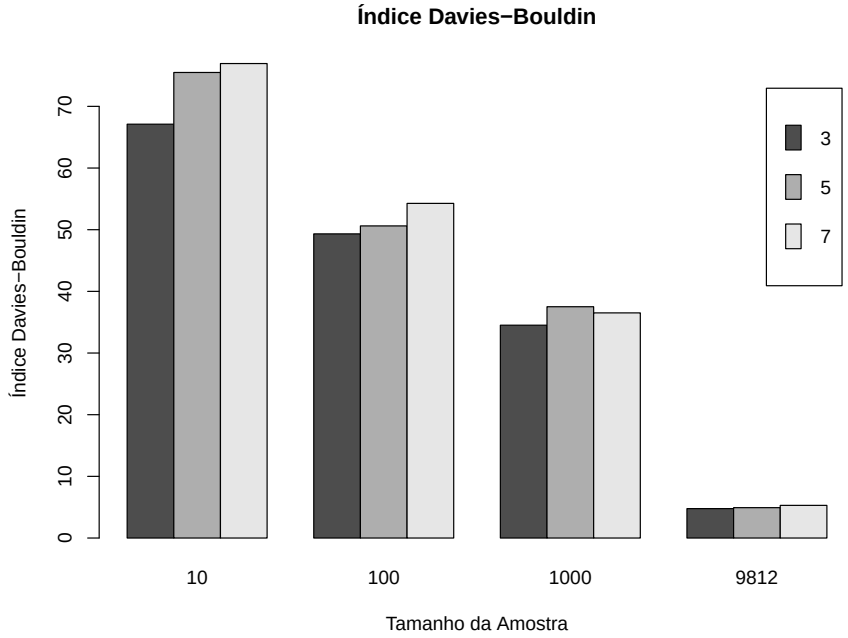

Figura 6: Índice Davies-Bouldin por K e tamanho de amostra

comportamento de outras medidas de distância, variar o tamanho das amostras e o número de grupos, assim como implementar novos métodos de redução de dados.

6 Cronograma de Execução

Para explorar a metodologia utilizada as atividades a serem realizadas foram definidas e detalhadas segundo seu período de execução, conforme a Tabela 5.

1. Finalização da implementação do método de amostragem baseado em fractais;
2. Realização de experimentos adicionais, comparando diferentes tipos de amostragem, conjuntos de dados e parâmetros;
3. Redação de artigo para publicação dos resultados;
4. Preparação da dissertação;
5. Defesa da dissertação.

Uberlândia, 14 de Dezembro de 2015.

Assinatura do Orientador:

Assinatura do Aluno:

Atividades	2016						2017	
	Jul.	Ago.	Set.	Out.	Nov.	Dez.	Jan.	Fev.
1								
2								
3								
4								
5								

Tabela 5: Cronograma de execução

Referências

[emb 2016] (2016). Embrapa informática agropecuária.

[glo 2016] (2016). Global land cover facility.

[Ding et al. 2015] Ding, R., Wang, Q., Dang, Y., Fu, Q., Zhang, H., and Zhang, D. (2015). YADING: Fast Clustering of Large-Scale Time Series Data. *VLDB Endowment*, 8(5):473–484.

[García et al. 2015] García, S., Luengo, J., and Herrera, F. (2015). *Data Preprocessing in Data Mining*, chapter Data Reduction, pages 147–162. Springer International Publishing.

[Gonçalves et al. 2014] Gonçalves, R., Zullo, J., Amaral, B. F. d., Coltri, P. P., Sousa, E. P. M. d., and Romani, L. A. S. (2014). Land use temporal analysis through clustering techniques on satellite image time series. In *Geoscience and Remote Sensing Symposium (IGARSS), 2014 IEEE International*, pages 2173–2176. IEEE.

[Guha et al. 1998] Guha, S., Rastogi, R., and Shim, K. (1998). Cure: an efficient clustering algorithm for large databases. In *ACM SIGMOD Record*, volume 27, pages 73–84. ACM.

[Han et al. 2011] Han, J., Pei, J., and Kamber, M. (2011). *Data Mining: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science.

[Kaufman and Rousseeuw 1990] Kaufman, L. and Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Series in Probability and Statistics. Wiley.

[Morettin and de Castro Toloi 2006] Morettin, P. and de Castro Toloi, C. (2006). *Análise de séries temporais*. ABE - Projeto Fisher. Edgard Blucher.

[Petitjean et al. 2014] Petitjean, F., Forestier, G., Webb, G. I., Nicholson, A. E., Chen, Y., and Keogh, E. (2014). Dynamic time warping averaging of time series allows faster and more accurate classification. In *Data Mining (ICDM), 2014 IEEE International Conference on*, pages 470–479. IEEE.

[Schroeder 1991] Schroeder, M. (1991). *Fractals, Chaos, Power Laws: Minutes From an Infinite Paradise*. W. H. Freeman.

[Tan et al. 2009] Tan, P., Steinbach, M., Kumar, V., and FERNANDES, A. (2009). *Introdução ao datamining: mineração de dados*. Ciencia Moderna.

[Traina-Jr. et al. 2010] Traina-Jr., C., Traina, A. J. M., Wu, L., and Faloutsos, C. (2010). Fast feature selection using fractal dimension. *Journal of Information and Data Management (JIDM)*, 1(1):3–16.

[Zaki and Meira 2014] Zaki, M. and Meira, W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.