SBBD 2021 ONLINE

Generalização de Mineração de Sequências Restritas no Espaço e no Tempo

Antonio Castro[1], Heraldo Borges[1],
Ricardo Campisano[1], Esther Pacitti[2,3], Fabio Porto[4],
Rafaelli Coutinho[1], Eduardo Ogasawara[1]

CEFET/RJ

UNIVERSITE MONTPELLIER

informatics mathematics Inria

LNCC

Introdução

- Evolução tecnológica
  - Popularização dispositivos digitais com sensores e GPS
  - Viabiliza grandes conjuntos de dados espaço-temporais
- Dados espaço-temporais → Importante para diferentes domínios
  - Oportunidade de extrair padrões interessantes
- Eventos com baixa frequência
- Buscar intervalo de tempo e posições onde os eventos são frequentes

Introdução – Exemplo Motivacional

- Engarrafamentos no Rio de Janeiro
  - Padrão: Proximidade do início/fim do expediente gera engarrafamentos
  - Alto suporte

- Em dias de jogo no Maracanã
  - Engarrafamento ao fim do jogo
  - Baixo suporte

- **Objetivo:** sequências de eventos, o conjunto de posições e o intervalo de tempo
  - Importante para o planejamento e gestão da cidade

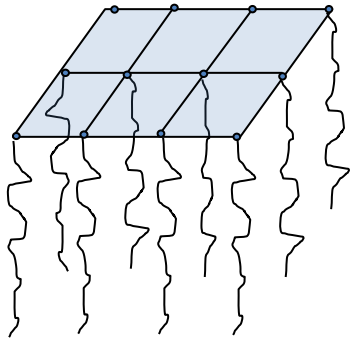
Trabalhos Relacionados

- Diferentes métodos:
  - Mineração = apenas tempo
    - Ben Chaabene *et al.* (2021), Xue *et al.* (2016).
  - Mineração + agrupamento = tempo + espaço
    - Koseoglu *et al.* (2020), Sunitha and Rama Mohan Reddy (2016)
- Restrições para espaço e tempo:
  - Suporte global = suporte válido para todo o conjunto de dados
    - Zhang *et al.* (2018), Batu *et al.* (2017)
  - Suporte local = janelas pré-definidas de tempo, espaço ou ambos
    - Chen *et al.* (2020), Koseoglu *et al.* (2020)

Proposta

Objetivo: Buscar sequências frequentes no tempo que ocorrem em grupos espaciais

- Sem restrições prévias para espaço e tempo



- Limites de densidade estabelecidos pelo usuário:
  - Frequência mínima em um intervalo de tempo $= \gamma$
  - Distância máxima dos elementos de um grupo $= \sigma$
  - Mínimo de posições dentro de um grupo $= \beta$
- Capaz de encontrar diferentes tamanhos de sequências, intervalos de tempo e regiões do espaço

Contribuições

- Generalização de Campisano *et al.* (2018)
  - Campisano *et al.* (2018) considera o espaço de forma linear

- Frequentes no tempo → próximas no espaço (três dimensões)

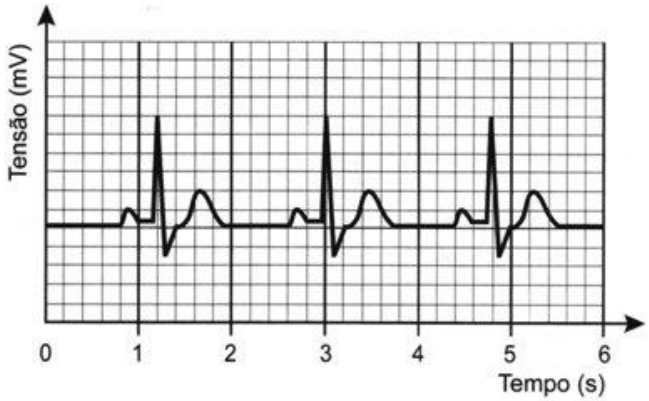
Conceitos Básicos

- **Sequência com marcação de tempo (TS)**

  - Uma sequência ordenada de observações obtidas por meio de medições repetidas ao longo do tempo

$$t = < v_1, v_2, \ldots, v_n >$$

  - Exemplo: eletrocardiograma



Conceitos Básicos

- **Subsequência**
  - Uma amostra contínua de uma TS, com tamanho e início definidos

$$sub_{m,p}(t)$$

  - Uma **sequência** faz parte de uma TS se existir uma posição a partir da qual seus itens sejam iguais, ou seja, se ela for subsequência da TS

$$s = < w_1, w_2, \ldots, w_k >, \exists q \mid s = sub_{k,q}(t) ,$$

    onde $|s| = k$

  - Exemplo: TS $t_1 = < B, A, B >$; $sub_{2,1}(t_1) = < B, A >$

	t_1
v_1	B
v_2	A
v_3	B

Conceitos Básicos

- **Posição:** Trio com as coordenadas de um ponto no sistema cartesiano

$$p = (x, y, z)$$

- **Sequência com marcação de tempo e espaço (STS)**
  - Par com uma posição e uma TS associada a esta posição

$$st = (p, t)$$

  - **Conjunto de dados** D: Conjunto de STS
  - Exemplo: $st_1 = (p_1, t_1)$, $st_2 = (p_2, t_2)$, e $st_3 = (p_3, t_3)$

	t_1	t_2	t_3
v_1	B	A	C
v_2	A	C	A
v_3	B	C	C

Conceitos Básicos

- **Suporte:** número de marcações de tempo no conjunto de dados STS nas quais a sequência ocorre

$$sup(s, D) = |Q|, \forall q \in Q, \exists st_i \in D \mid s = sub_{|s|,q}(st_i.t),$$

onde Q = conjunto de marcações de tempo em que s ocorre em D

- **Frequência:** divisão do suporte de uma sequência em um conjunto de dados STS pelo tamanho da TS no conjunto de dados

$$freq(s, D) = \frac{sup(s, D)}{|st.t|}, st \in D$$

  - Uma sequência é **frequente** se sua frequência no conjunto de dados STS for maior ou igual a frequência mínima definida pelo usuário

$$freq(s, D) \geq \gamma$$

- **Período de tempo (período):** definido por uma marcação de tempo inicial e uma final

$$r = (r_s, r_e)$$

Conceitos Básicos

Exemplo:

- $s = < A, C >$
- $sup(s, D) = 2$ (ocorre em v_1 e v_2)
- $freq(s, D) = \frac{2}{3}$

	t_1	t_2	t_3
v_1	B	$\boldsymbol{A}$	C
v_2	A	$\boldsymbol{C}$	$\boldsymbol{A}$
v_3	B	C	$\boldsymbol{C}$

Conceitos Básicos e *Ranged Group* (RG)

- **Grupo de posições espaciais (grupo):** conjunto de posições onde seus elementos devem estar a uma distancia máxima definida pelo usuário (σ) de ao menos um outro elemento do mesmo grupo

$$g \mid \forall p \in g, \exists q \in g \mid dist(p,q) \leq \sigma$$

- ***Ranged Group* (RG):** trio (s, r, g)
  - s é uma sequência
  - r é um período de tempo
  - g é um grupo de posições espaciais

Ranged Group (RG)

- **Ocorrências** de uma sequência em um RG: número de vezes em que a sequência ocorre no intervalo de tempo, e conjunto de STS do grupo, referentes ao RG

$$occur(s, r, g)$$

- **Suporte** de uma sequência em um RG: número de marcações de tempo em que a sequência começa no intervalo de tempo, e no conjunto de STS do grupo, referentes ao RG

$$sup(s, r, g) = |Q|, \forall q \in Q, \exists st \in sts(g) \mid s = sub_{|s|,q}(st.t), r_s \leq q \leq r_e, |s| \leq r_e$$

- **Frequência** de uma sequência em um RG: divisão do suporte do RG pelo tamanho de do intervalo de tempo do RG

$$freq(s, r, g) = \frac{sup(s, r, g)}{|r|}$$

Ranged Group (RG)

Exemplo:

- $\sigma = 1$
- $RG(s, r, g)$:
  - $s = < C >$
  - $r = [1, 3]$,
  - $g = < p_2, p_3 >$

$occur(s, r, g) = 4$ $\quad sup(s, r, g) = 3$ $\quad freq(s, r, g) = \frac{3}{3}$

	t_1	t_2	t_3
v_1	B	A	C
v_2	A	C	A
v_3	B	C	C

Kernel Ranged-Group (KRG)

Um **Kernel Ranged-Group (KRG)** é um RG com as seguintes restrições:

1. A frequência deve ser maior que a mínima definida pelo usuário
$$freq(s, r, g) \geq \gamma$$
2. O tamanho do grupo deve ser maior que o mínimo definido pelo usuário
$$|g| \geq \beta$$
3. Período de tempo mínimo: reduzir o período de tempo mantém a frequência maior ou igual a mínima, mas o suporte diminui
$\forall r' \in PR \mid r' \subset r$ e $r'_s = r_s$, ambas as condições se aplicam:
   a) $freq(s, r', g) \geq \gamma$
   b) $sup(s, r', g) < sup(s, r, g)$
4. Grupo máximo e mínimo: aumentar o grupo não aumenta o número de ocorrências e diminuir o grupo reduz o número de ocorrências
   a) $\forall g' \in PG \mid g \subseteq g', occur(s, r, g') = occur(s, r, g)$
   b) $\forall g' \in PG \mid g' \subset g, occur(s, r, g') < occur(s, r, g)$

Kernel Ranged-Group (KRG)

Exemplo:

- $\gamma = 60\%,\ \beta = 2,\ \sigma = 1$
- $RG(s,r,g)$:
  - $s = C$
  - $r = [1,3]$
  - $g =< p_2, p_3 >$

$occur(s,r,g) = 4$

$sup(s,r,g) = 3$

$freq(s,r,g) = \frac{3}{3} = 1$

	t_1	t_2	t_3
v_1			C
v_2		C	
v_3		C	C
v_4			

Kernel Ranged-Group (KRG)

Exemplo:

- $\gamma = 60\%,\ \beta = 2,\ \sigma = 1$
- $RG(s,r,g)$:
  - $s = C$
  - $r = [1,3]$
  - $g =< p_2, p_3 >$

$occur(s,r,g) = 4$

$sup(s,r,g) = 3$

$freq(s,r,g) = \frac{3}{3} = 1$

	t_1	t_2	t_3
v_1			C
v_2		C	
v_3		C	C
v_4			

Diminuir r para $[1,2]$
faz $\sup(s,r,g) = 2$

	t_1	t_2	t_3
v_1			C
v_2		C	
v_3		C	C
v_4			

Kernel Ranged-Group (KRG)

Exemplo:

- $\gamma = 60\%,\ \beta = 2,\ \sigma = 1$
- $RG(s,r,g)$:
  - $s = C$
  - $r = [1,3]$
  - $g = < p_2, p_3 >$

$occur(s,r,g) = 4$

$sup(s,r,g) = 3$

$freq(s,r,g) = \frac{3}{3} = 1$

	t_1	t_2	t_3
v_1			C
v_2		C	
v_3		C	C
v_4			

Incluir p_1 no grupo mantém o mesmo número de ocorrências

	t_1	t_2	t_3
v_1			C
v_2		C	
v_3		C	C
v_4			

Kernel Ranged-Group (KRG)

Exemplo:

- $\gamma = 60\%,\ \beta = 2,\ \sigma = 1$
- $RG(s,r,g)$:
  - $s = C$
  - $r = [1,3]$
  - $g =< p_2, p_3 >$

$occur(s,r,g) = 4$

$sup(s,r,g) = 3$

$freq(s,r,g) = \frac{3}{3} = 1$

	t_1	t_2	t_3
v_1			C
v_2		C	
v_3		C	C
v_4			

Remover p_2 do grupo faz $occur(s,r,g) = 2$

	t_1	t_2	t_3
v_1			C
v_2		C	
v_3		C	C
v_4			

Solid Ranged-Group (SRG)

Objetivo: Padrões restritos no espaço e no tempo, construídos a partir de KRG ⇒ Encontrar os **Solid Ranged-Groups (SRG)**

Um **SRG** é um RG com algumas restrições além das referentes aos KRG:

- O período de tempo é mínimo e máximo

  Aumentar o período de tempo diminui a frequência para abaixo da mínima e/ou o suporte se mantém

  Diminuir o período de tempo reduz o suporte

  i. $\forall r' \in PR \mid r \subseteq r'$, temos a) ou b) ou ambos:

  a) $freq(s, r', g) < \gamma$

  b) $sup(s, r', g) = sup(s, r, g)$

  ii. $\forall r' \in PR \mid r' \subset r,\ sup(s, r', g) < sup(s, r, g)$

Solid Ranged-Group (SRG)

Exemplo:

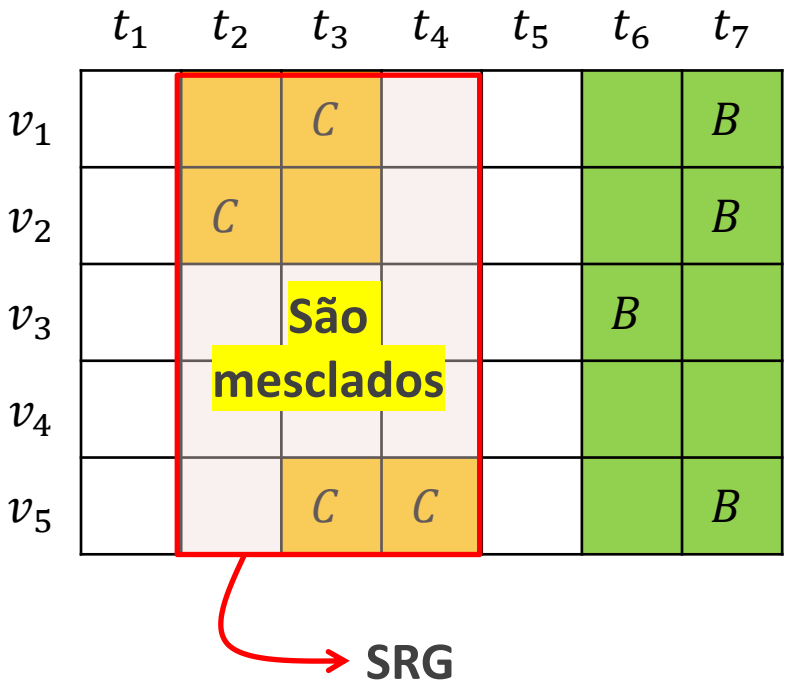
- $\gamma = 60\%,\ \beta = 2,\ \sigma = 1$

	t_1	t_2	t_3	t_4	t_5	t_6	t_7
v_1			C				B
v_2		C					B
v_3						B	
v_4							
v_5			C	C			B

Solid Ranged-Group (SRG)

Exemplo:

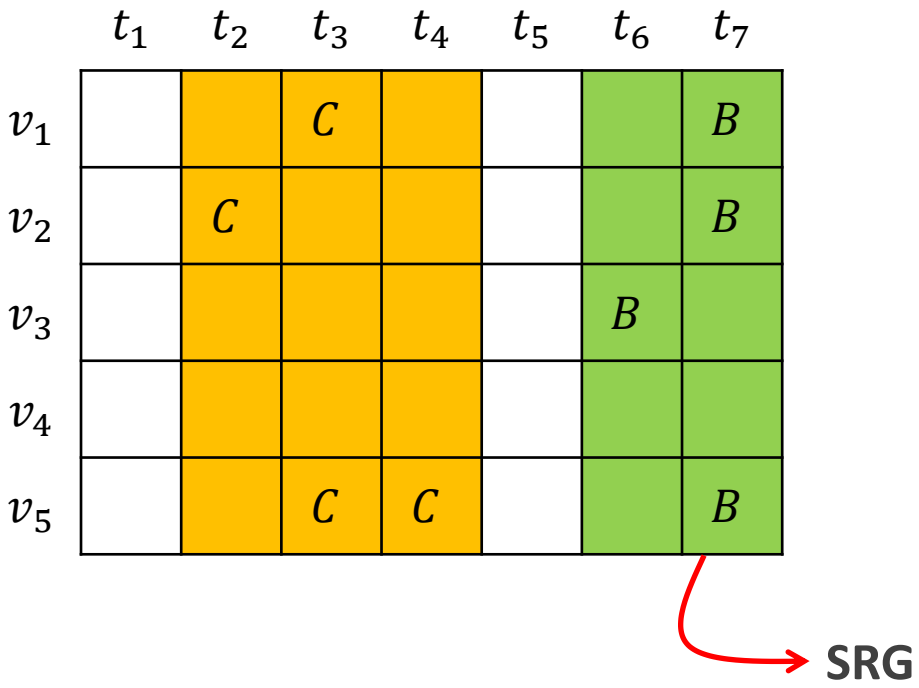
- $\gamma = 60\%,\ \beta = 2,\ \sigma = 1$



Solid Ranged-Group (SRG)

Exemplo:

- $\gamma = 60\%,\ \beta = 2,\ \sigma = 1$



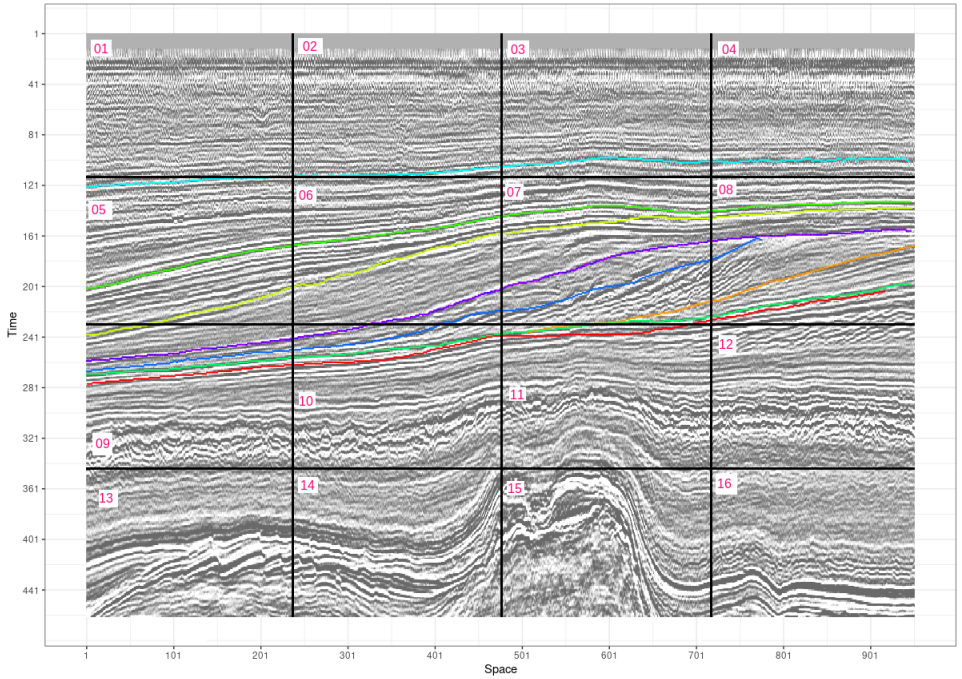
Avaliação Experimental

- Algoritmo capaz de produzir os SRGs
- Análise de sensibilidade do algoritmo

Os experimentos foram implementados em R e executados em um computador com processador com 16 núcleos, 128GB de RAM e Ubuntu 20.04 LTS

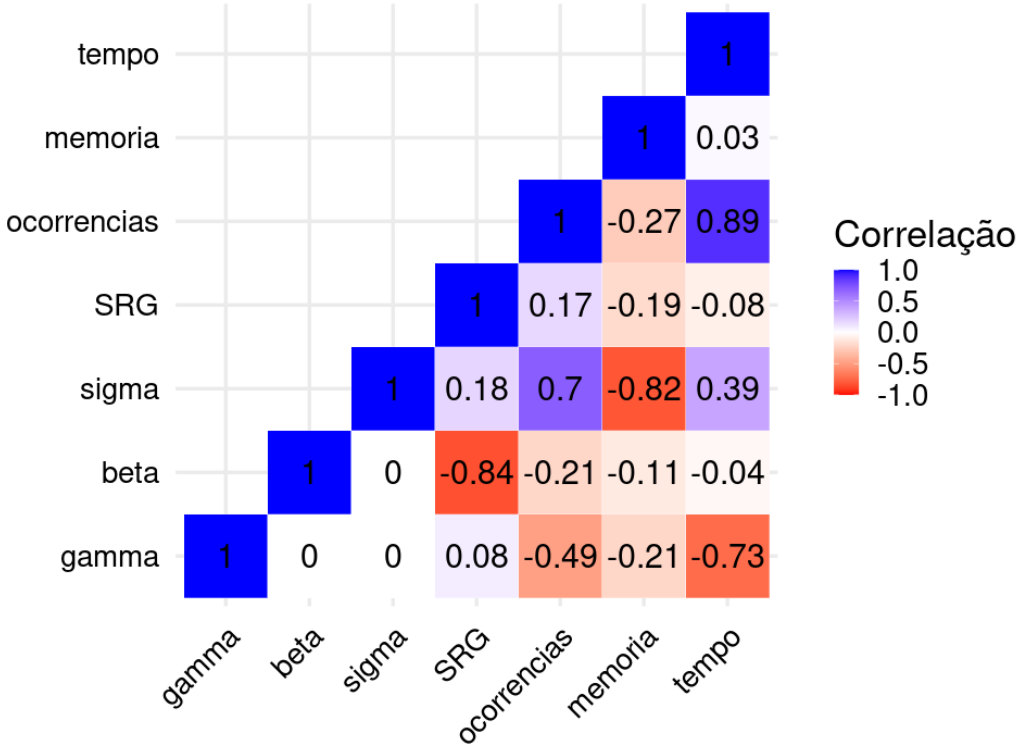
Conjunto de Dados

- Conjunto de Dados
  - $D \rightarrow$ *inline* T401 do F3 Block
- 951 sequências com marcação de tempo e espaço com 462 observações discretizado
- Alfabeto de tamanho 25
- Dividido em 16 quadrantes organizados de maneira retangular (4 x 4)



Análise de Sensibilidade

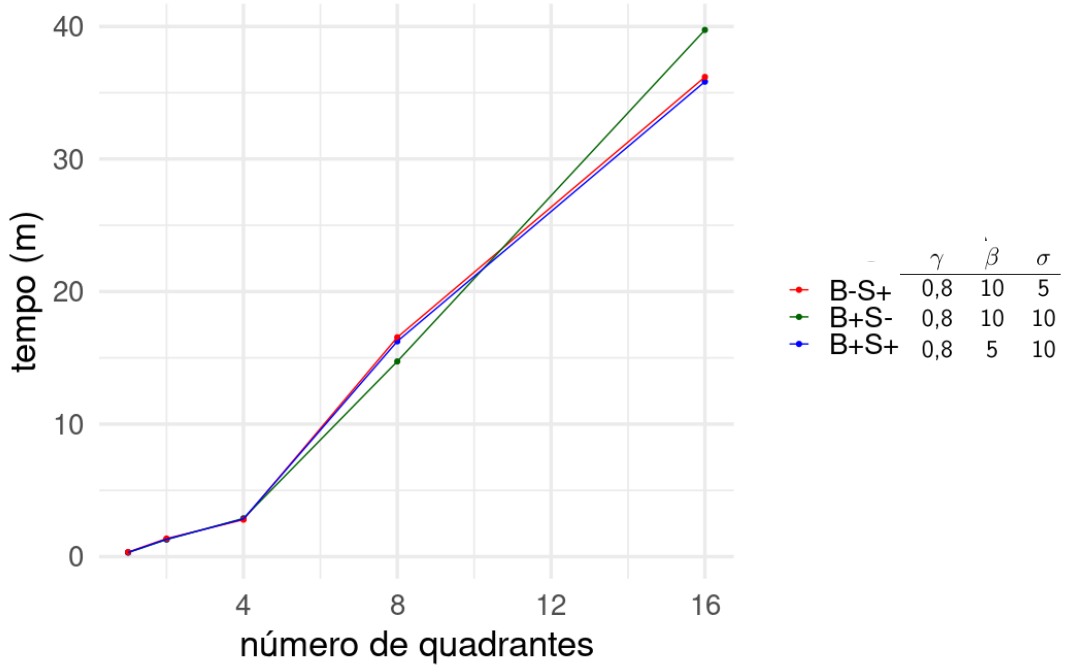
- Correlação entre os parâmetros de entrada (γ, β e σ) e as informações de saída:
  - número de SRG
  - número de ocorrências
  - uso de memória
  - tempo de execução



$\gamma = \{0{,}6;\ 0{,}8;\ 1{,}0\}, \beta = [5{,}10]\ e\ \sigma = [5{,}10]$

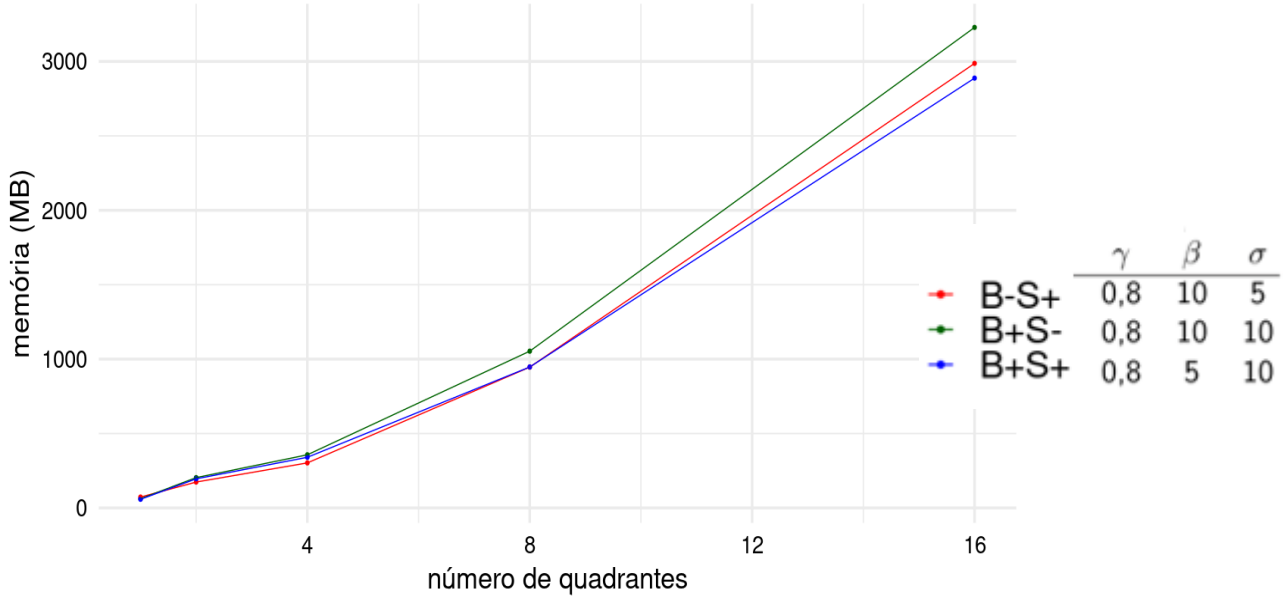
Análise de Sensibilidade

- Tempo de execução usando diferentes configurações e tamanhos de conjunto de dados



Análise de Sensibilidade

- Uso máximo de memória para diferentes configurações e tamanhos de conjunto de dados



Conclusões

- Fundamentos importantes e as noções de grupo, RG, KRG e SRG foram introduzidos
- Primeiro algoritmo capaz de encontrar sequências restritas no espaço e no tempo que funciona com uma dimensão de tempo e três dimensões de espaço
- Experimentos com conjunto de dados sísmicos do mundo real
- Análise de sensibilidade do algoritmo ⇒
  - o que mais afeta seu funcionamento é o tamanho do conjunto de dados de entrada
- Trabalhos em andamento:
  - Comparação com uma abordagem intuitiva
  - Avaliação do algoritmo com dados de outros domínios

SBBD 2021 ONLINE

CEFET/RJ

Generalização de Mineração de Sequências Restritas no Espaço e no Tempo

Obrigado!

Eduardo Ogasawara

eogasawara@ieee.org

CAPES

CNPq Conselho Nacional de Desenvolvimento Científico e Tecnológico

FAPERJ