Motifs em Séries Espaço Temporais

Murillo G. Dutra[1]

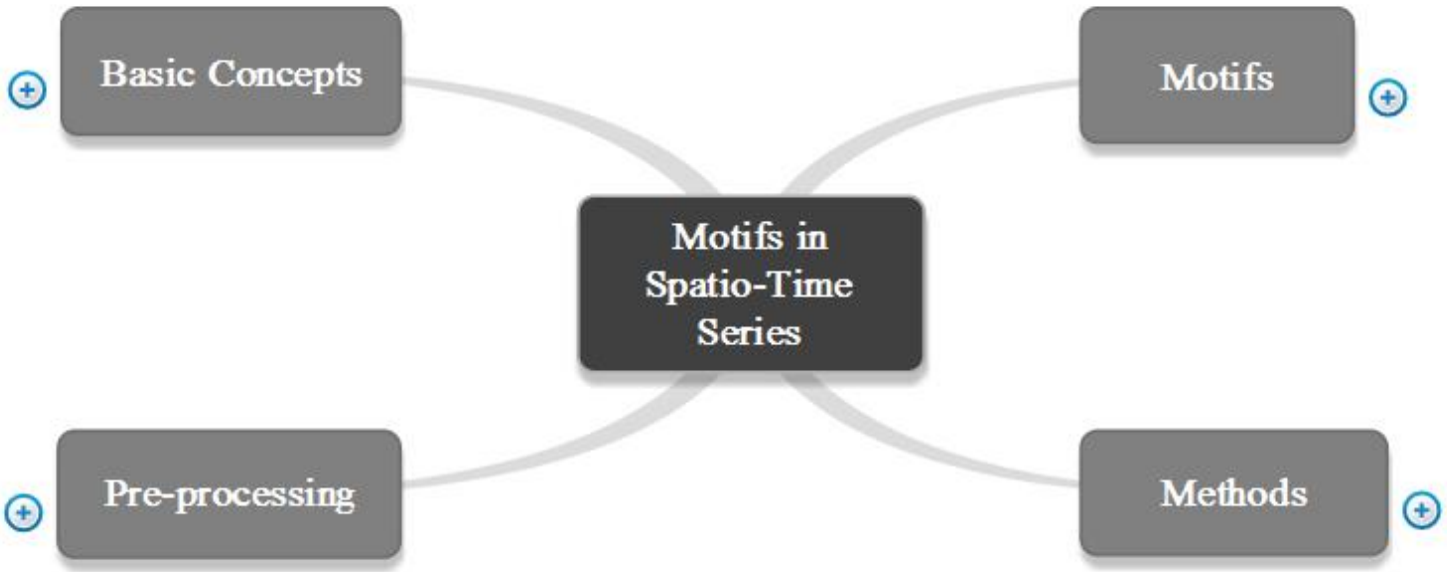
[1]Aluno de Mestrado do Programa de Pós-Graduação em Tecnologia (PPTEC)
CEFET/RJ
Orientador: Eduardo Ogasawara, DSc

20/05/2015

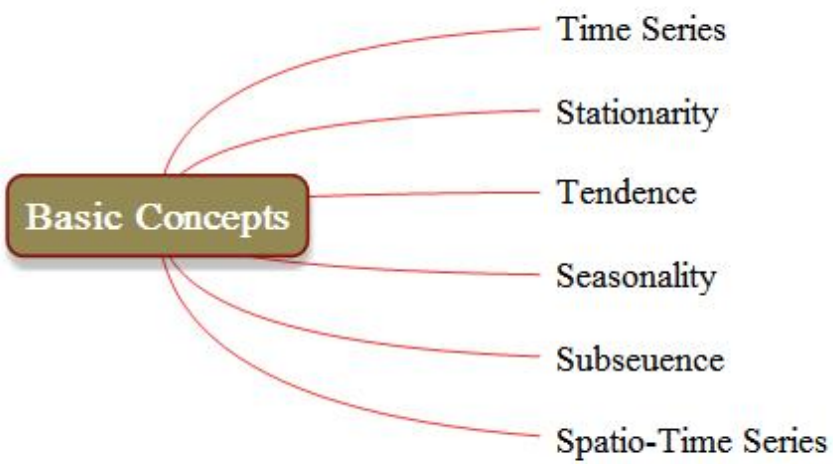
Motivação

- Necessidade de identificação de fenômenos de difícil observação sem uso computacional.
- Ampliação da extração de conhecimento de conjunto de dados que integram espaço e tempo demandam por modelagem ainda pouco exploradas.
- Amplo campo de aplicação.

Taxonomia

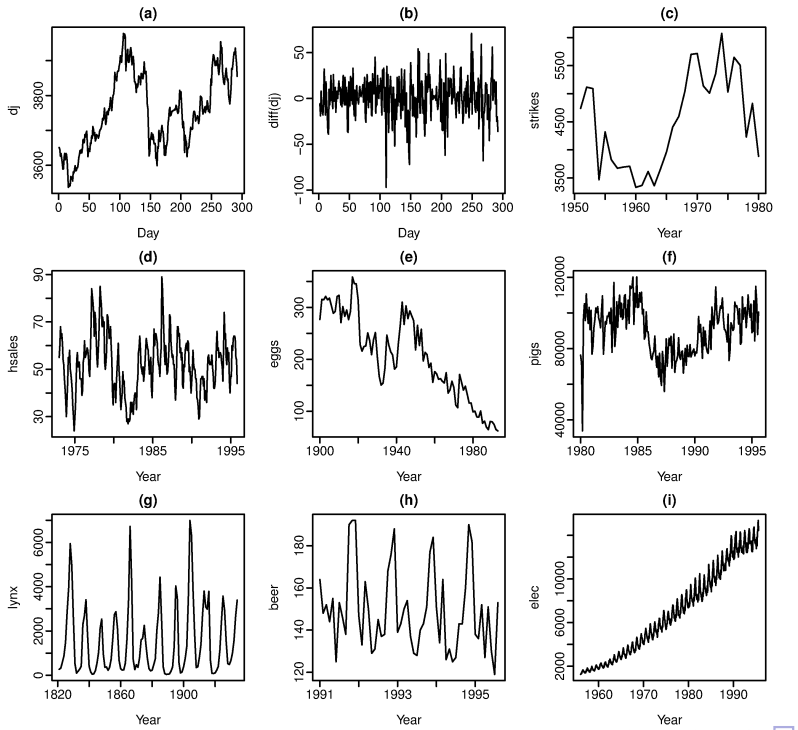

Basic Concepts
Motifs
Motifs in Spatio-Time Series
Pre-processing
Methods

Taxonomia - Basic Concepts



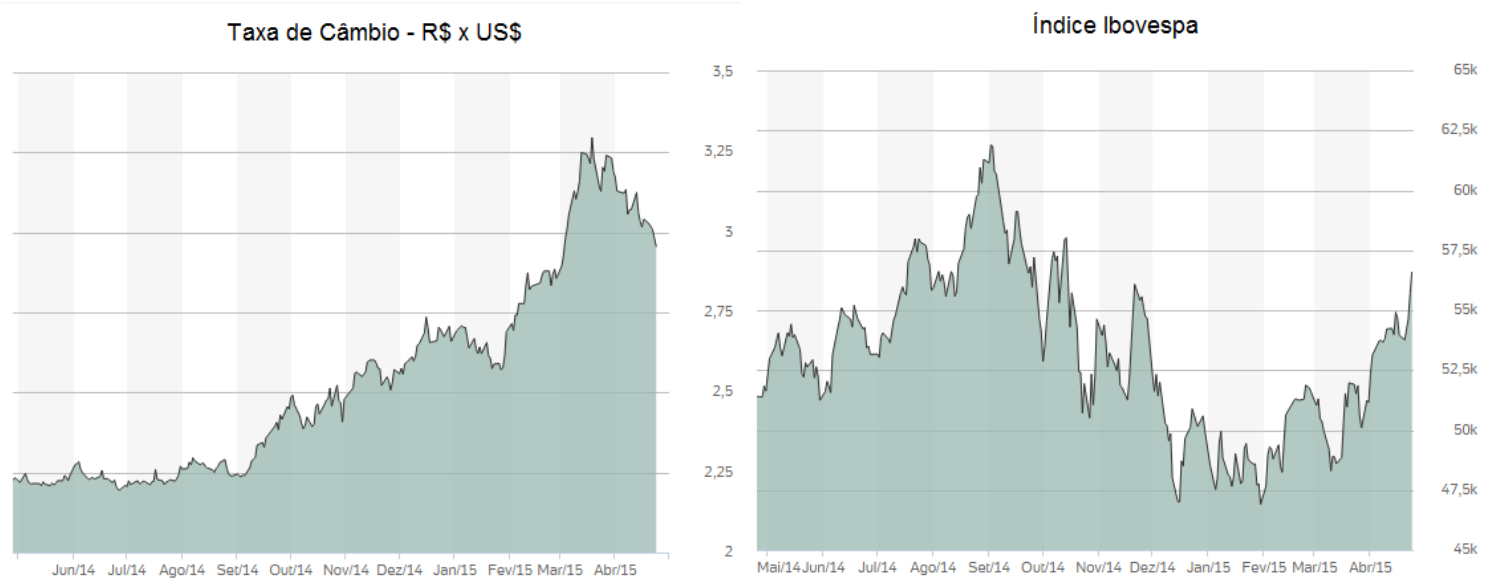
Séries Temporais

- Uma Série Temporal pode ser definida como uma sequência ordenada de valores ao longo de um período de tempo.
- Uma Série Temporal t é uma sequência ordenada de valores $\{v_1, ..., v_m\}$.
- $|t|$ é o número m de elementos em t.
- v_m é o valor mais recente de t.



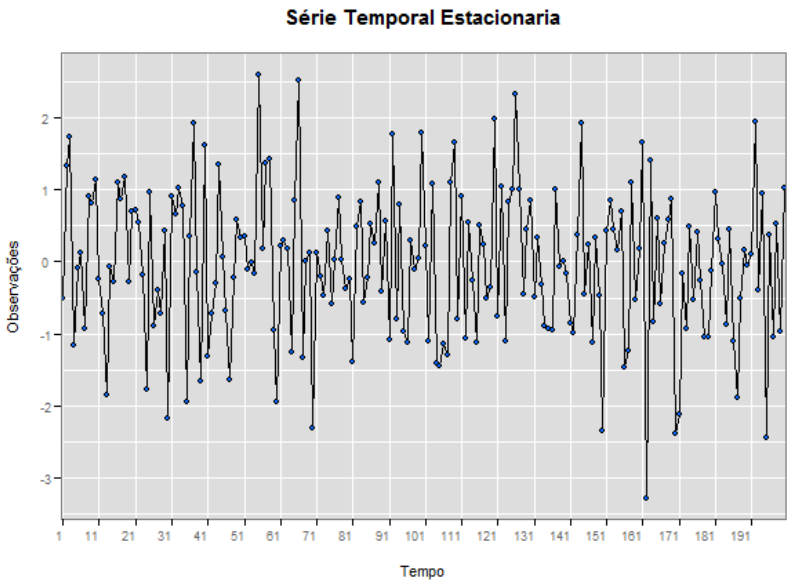
Discreta

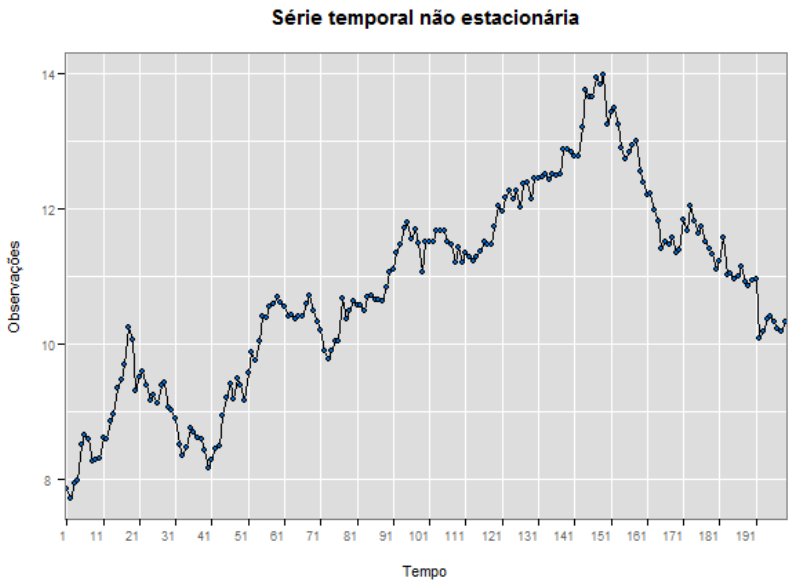
- Os valores são observados em intervalos de tempo definidos (Ano, Mês, dia, hora, etc). Exemplos: Índice Ibovespa, Taxa de Câmbio



Estacionaridade

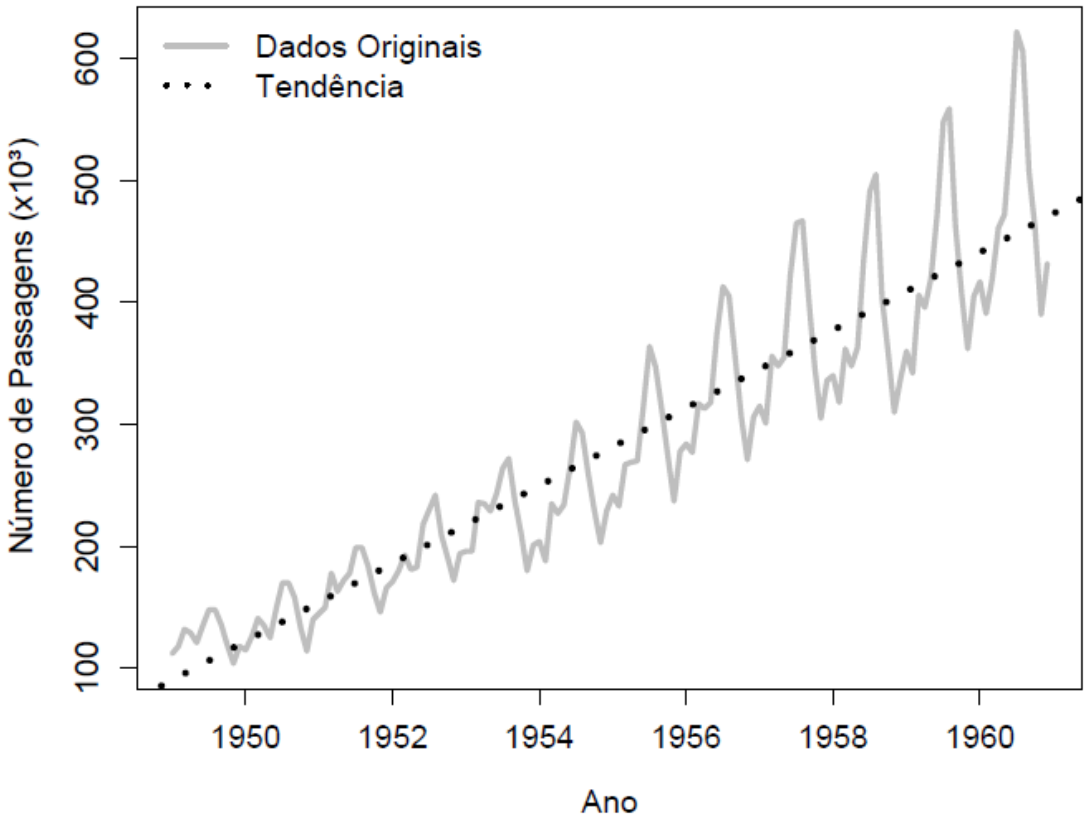
- Característica da série temporal em variar aleatoriamente ao redor de uma média constante.





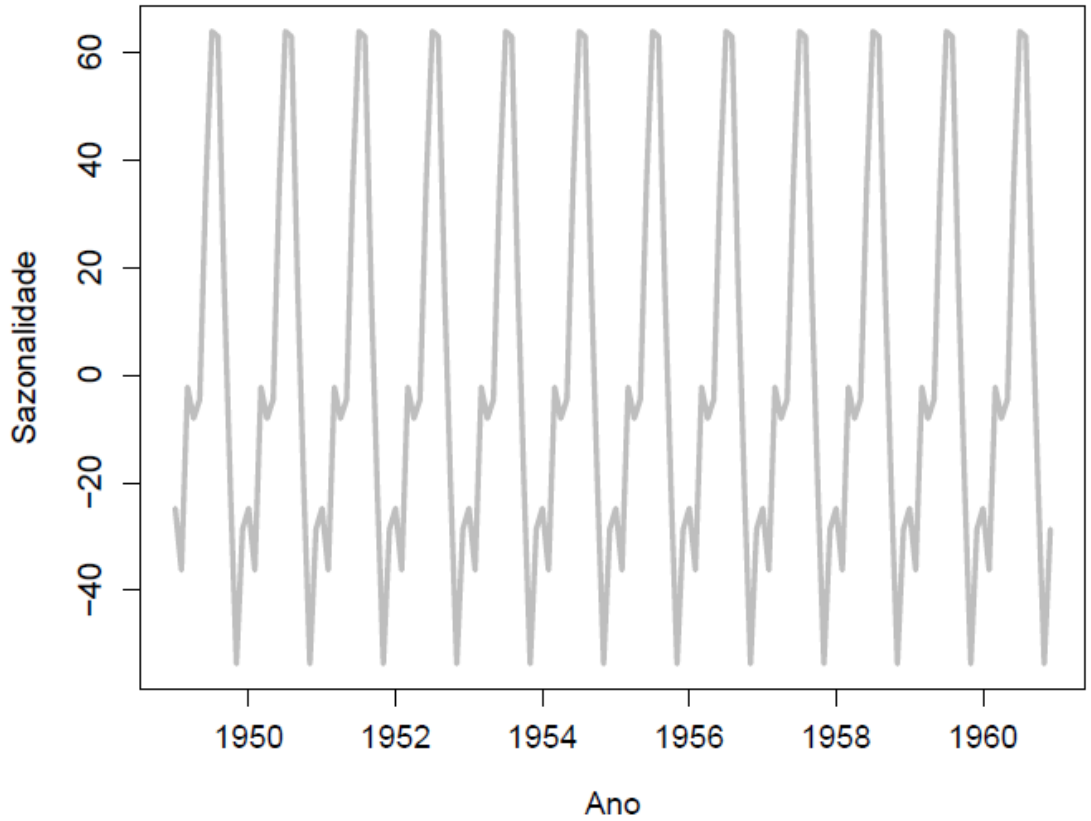
Tendência

- Característica esperada do comportamento de uma série temporal ao longo de um período de tempo.



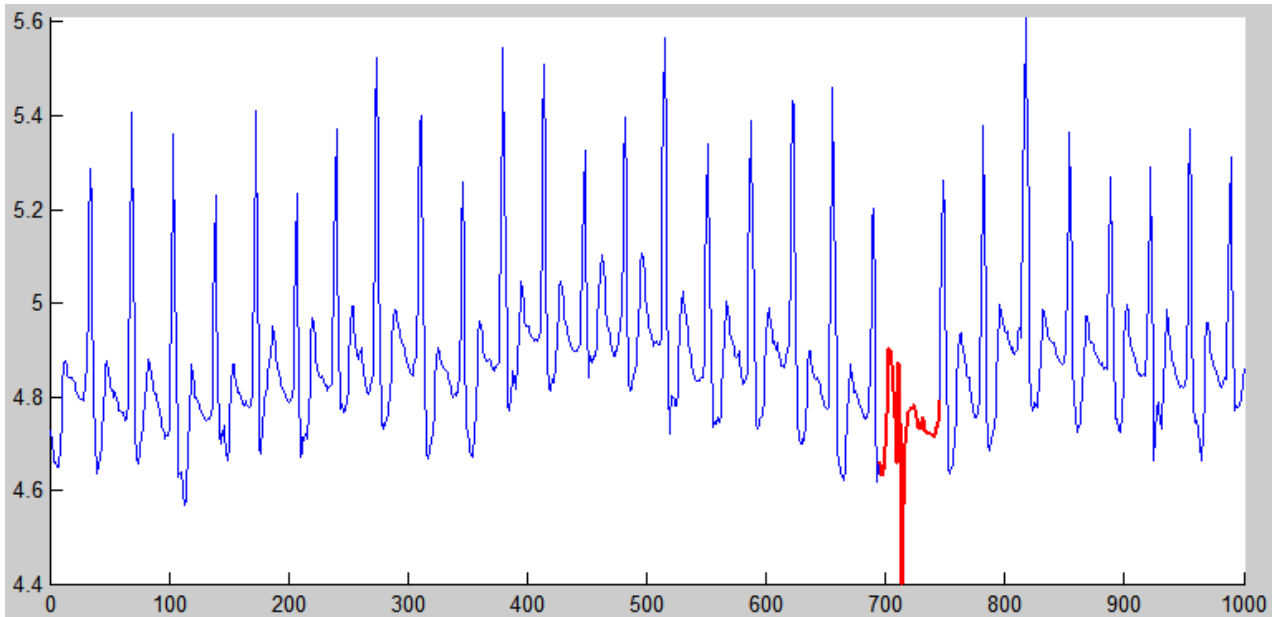
Sazonalidade

- São movimentos similares que ocorrem com periodicidade fixa.



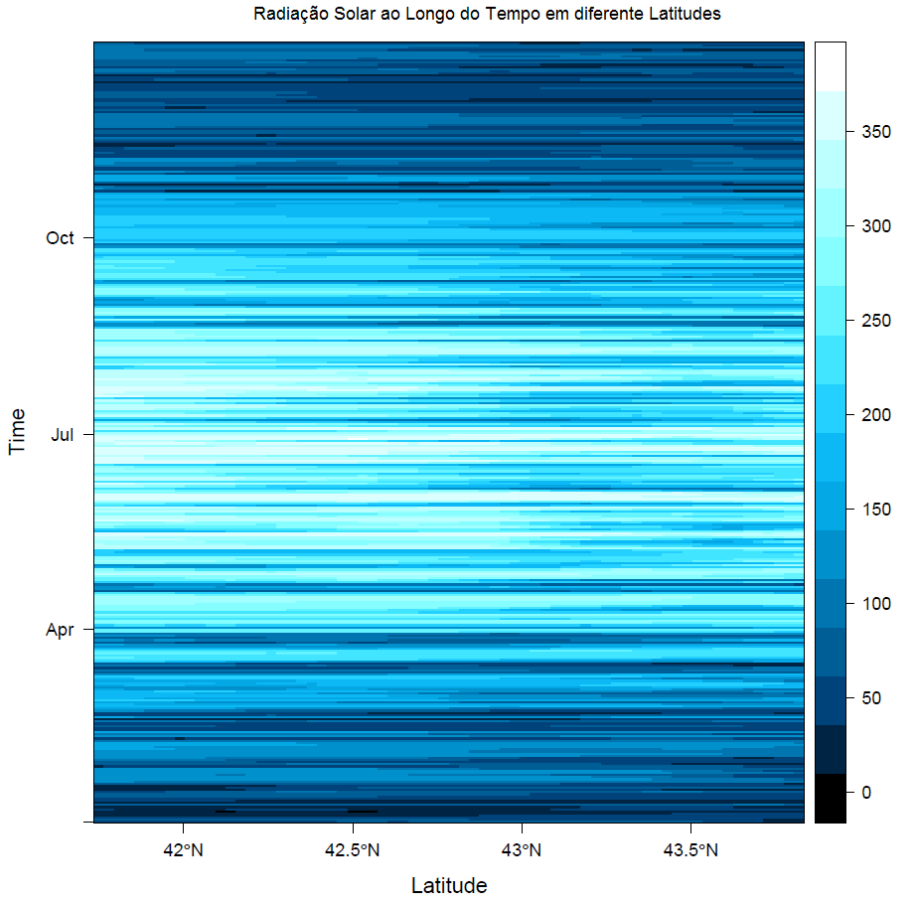
Subsequência

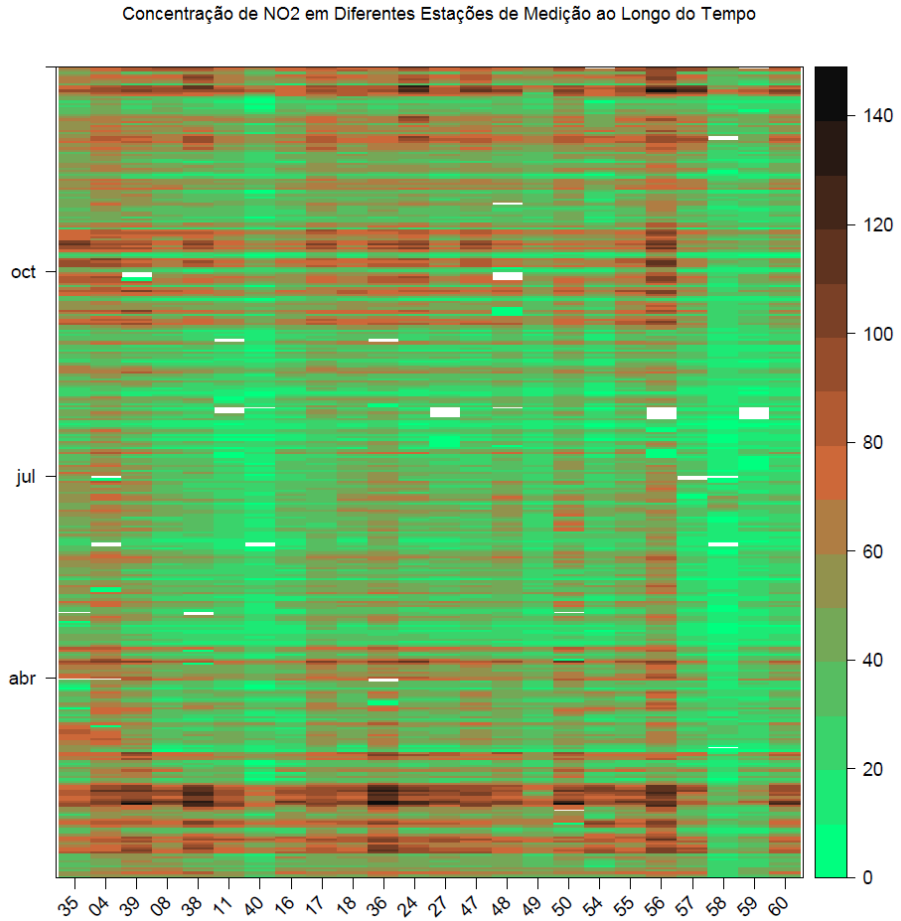
- Dada uma série temporal t de tamanho m, uma subsequencia C de t é uma amostra contínua de t de tamanho n, sendo n menor que m.



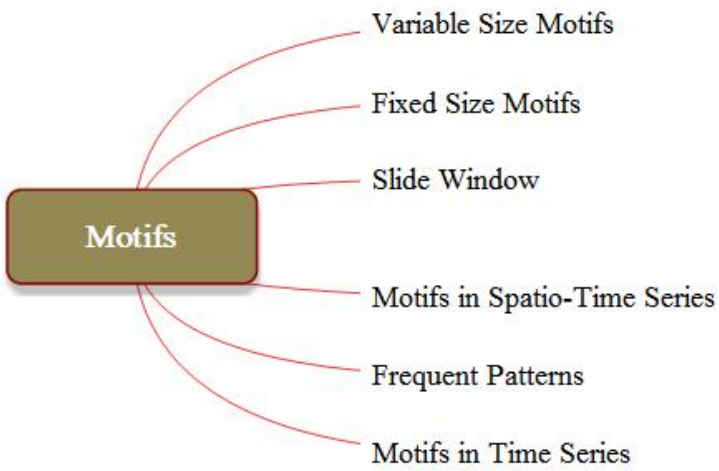
Série Espaço Temporal

- Uma série espaço temporal é uma série temporal com uma posição associada.



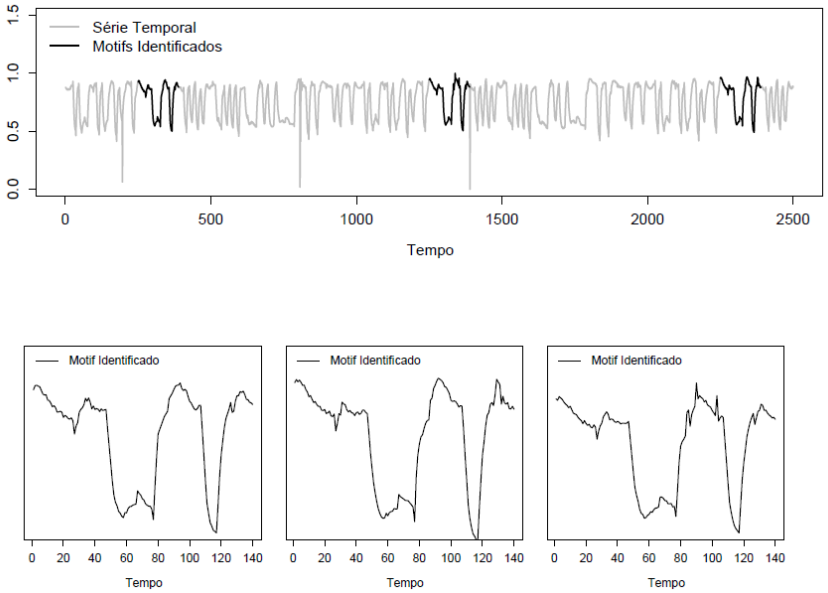


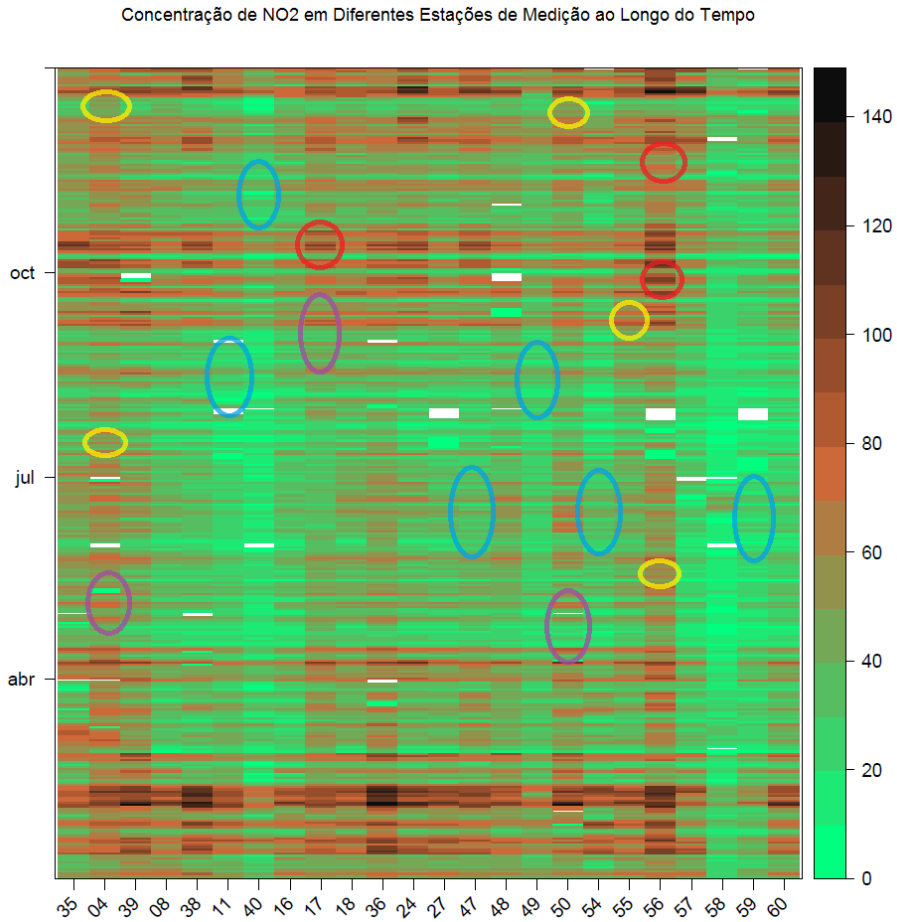
Taxonomia - Motifs



Motifs

- Padrões de subsequência frequentes previamente desconhecidos que se repetem em uma série temporal ou espaço temporal.

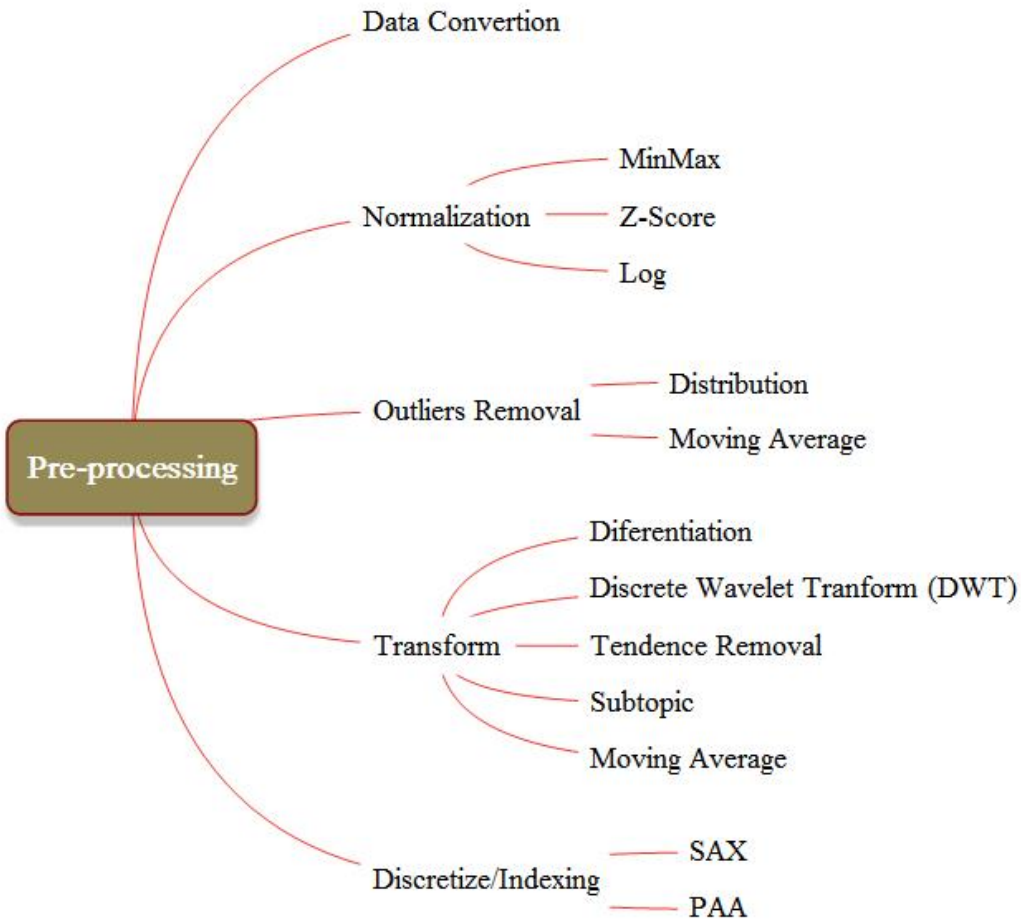




Identificação de Motifs

- **Pré-Processamento**
  - Conversão de Dados
  - Normalização
  - Remoção de Outliers
  - Transformação
  - Discretização/Indexação
- **Métodos**
  - Medidas de Distância
  - Algoritmos

Taxonomia - Pre-processing



Pré-processamento [Conversão]

- **Conversão dos Dados**
  - Eventualmente, o conjunto de dados não está num formato apto a ser trabalhado.
  - Nesse caso, é necessário converter do formato original do dado para um formato padrão, como o csv, para possibilitar a leitura dos dados em ferramentas de análise de dados.

Pré-processamento [Remoção de Outliers]

- **Remoção de Outliers:** Processo de limpeza dos dados para remoção de observações que possam prejudicar a análise dos dados.

Pré-processamento [Normalização]

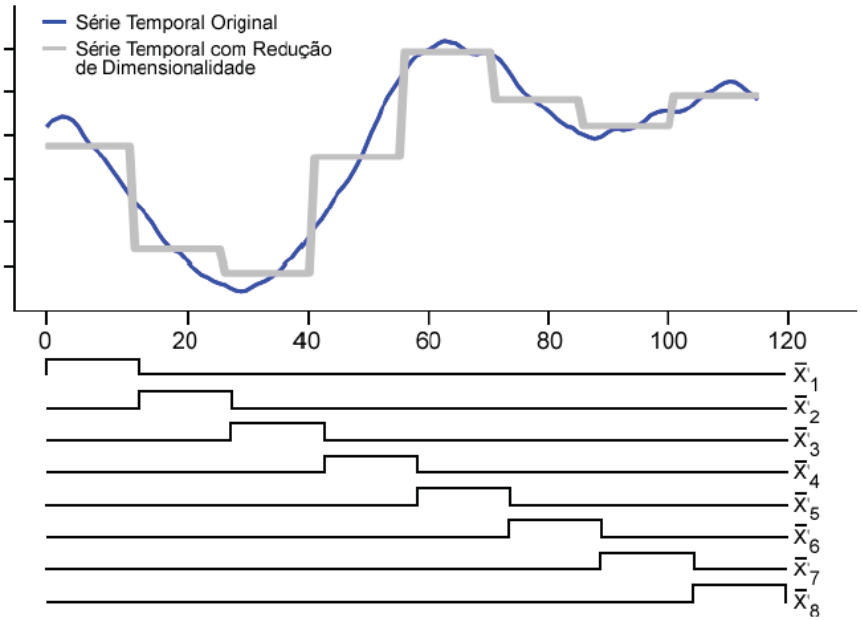
- **Normalização:** Para comparar séries temporais é fundamental que estejam em uma mesma base/escala de valores.
  Exemplos:
  - Normalização de Offset: Subtração de cada item da série pela média aritmética
  - Normalização de Amplitude: Similar a anterior, porém divide-se cada subtração pelo desvio padrão da série temporal
  - Normalização de Escala: Consiste em estabelecer a escala no intervalo [0,1]

Pré-processamento [Discretização]

- **Discretização:** É uma transformação dos valores reais numéricos de uma série temporal em valores discretos ou simbólicos. É importante para viabilizar a análise dos dados e aplicação de métodos para extração de conhecimento.

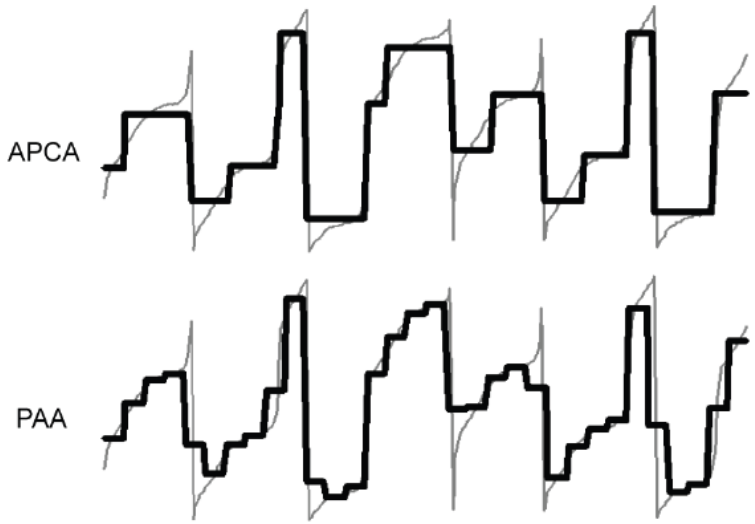
Pré-processamento [Discretização]

- **Piecewise Aggregate Approximation (PAA)**: Separação da série temporal em pequenos segmentos de mesmo tamanho. Para cada segmento é atribuido o valor da média aritmética.



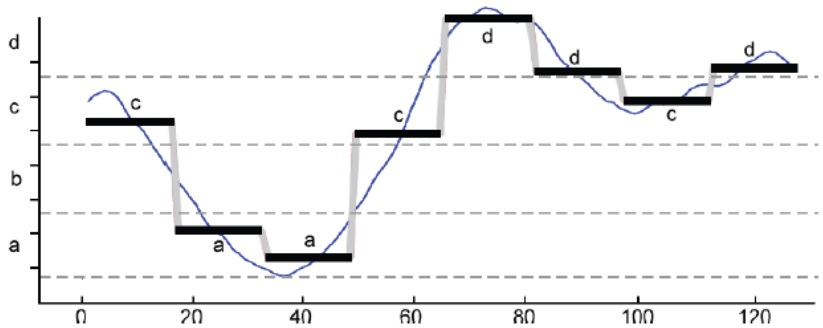
Pré-processamento [Discretização]

- **Adaptative Piecewise Constant Approximation (APCA)**: É uma adaptação do método anterior, a qual considera o comportamento da série para definir o tamanho dos segmentos.



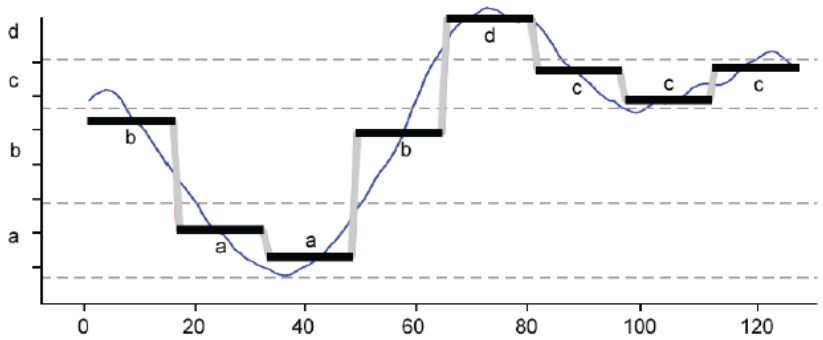
Pré-processamento [Discretização]

- **Particionamento Uniforme:** Converte uma série temporal em uma sequencia de símbolos. São obtidos os limites superior e inferior da série, desse forma, o intervalo entre os limites superior e inferior é dividido em um conjunto de regiões de mesmo tamanho. Cada uma dessas regiões recebe uma etiqueta de uma letra do alfabeto.



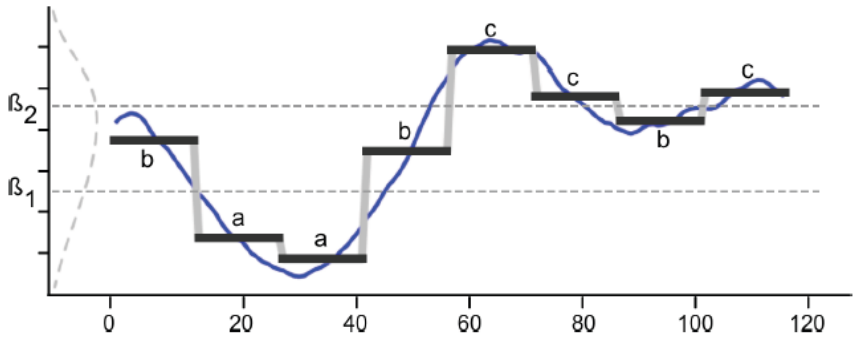
Pré-processamento [Discretização]

- **Particionamento por Entropia Máxima:** Nesse método, leva-se em consideração que determinados segmentos são mais ricos de informação que outros. Dessa forma, a divisão do segmento será em função da riqueza da informação de cada região.

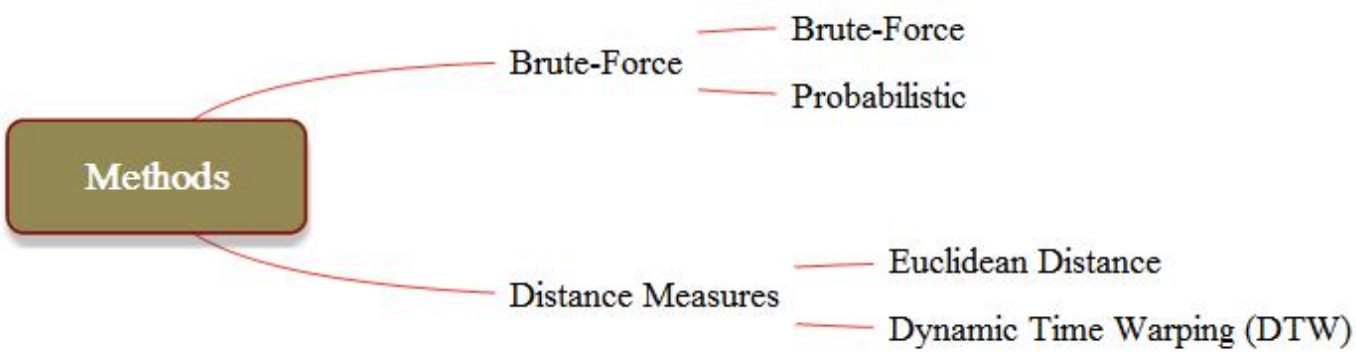


Pré-processamento [Discretização]

- **Symbolic Aggregate approXimation (SAX):** As observações das subsequencias tendem a ser normalmente distribuidas. Dessa forma, o espaço de discretização é feita sobre o sino Gaussinado em intervalos diferentes com mesma probabilidade.



Taxonomia - Métodos



Métodos

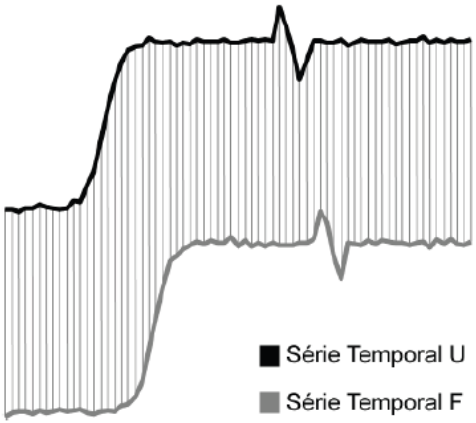
- **Similaridade em Séries Temporais**
  - São métodos para comparação de séries temporais e/ou subsequências.
  - Com isso é possível identificar padrões frequentes nas séries temporais dentro da mesma série ou em séries diferentes.

Métodos[Medidas de Distância]

- **Medidas de Distância**
  - Distância Euclideana
  - *DynamicTimeWarping*(DTW):

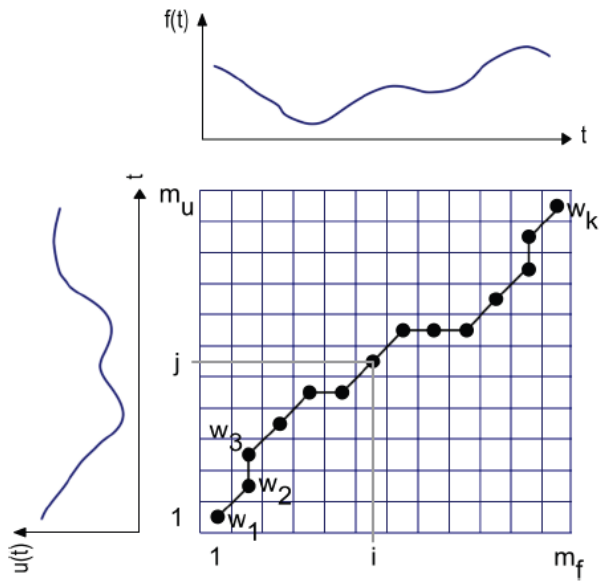
Métodos[Medidas de Distância]

- - **Distância Euclideana:** Determina o comprimento da linha reta entre dois pontos das séries



Métodos[Medidas de Distância]

- **Dynamic Time Warping(DTW):** Usado para comparação de séries com defasagem no tempo e de tamanhos diferentes. Na determinação da DTW deve-se alinhar ambas as séries temporais por meio da construção de uma matriz na qual os elementos da matriz são as distâncias euclideanas para determinar a distância entre dois pontos.



Matriz de Distância

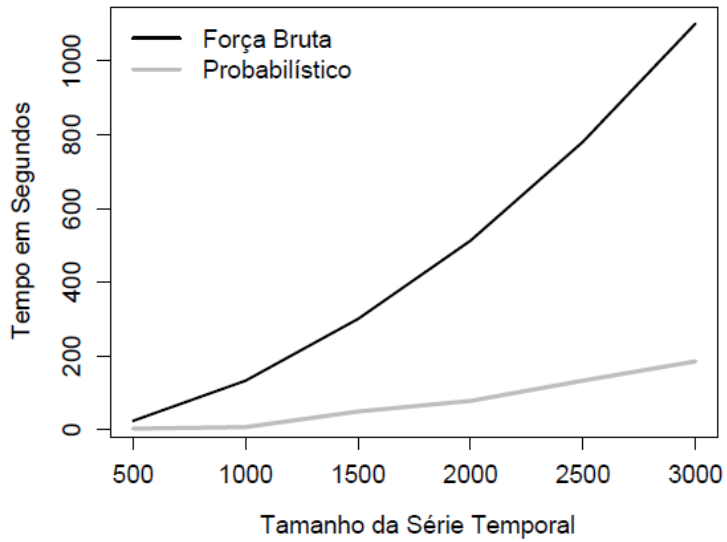
0	0	1	4	1	0	0
1	1	0	1	0	1	1
4	4	1	0	1	4	4
4	4	1	0	1	4	4
1	1	0	1	0	1	1
0	0	1	4	1	0	0
1	1	0	1	0	1	1
0	0	1	4	1	0	0

Matriz de Distância Cumulativa

11	11	4	6	2	1	1
11	11	3	2	1	2	3
10	10	3	1	2	6	10
6	6	2	1	2	6	6
2	2	1	2	2	2	2
1	1	1	4	2	1	1
1	1	0	1	1	2	3
0	0	1	5	6	6	6

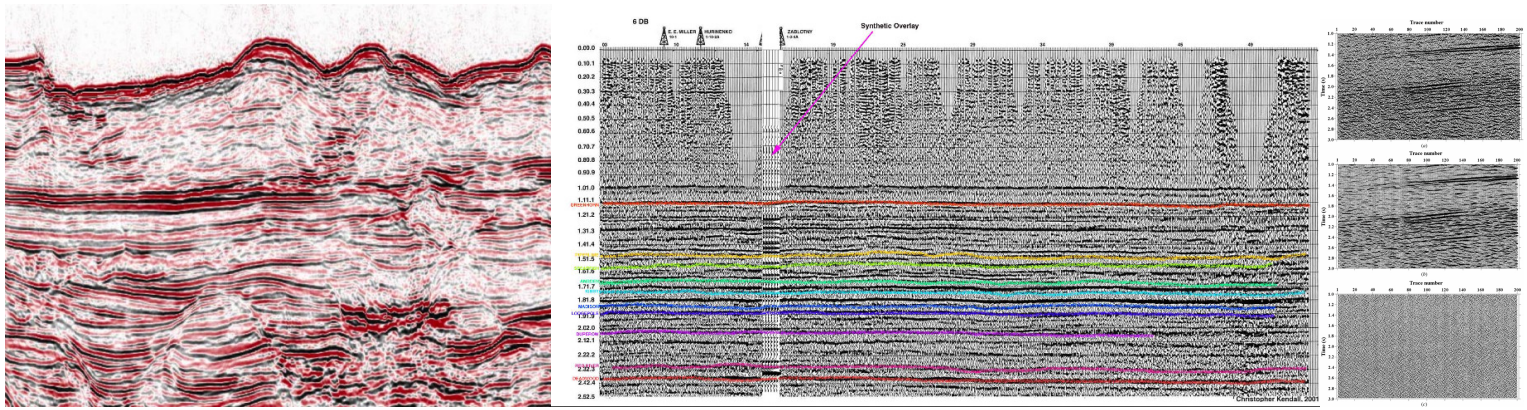
Métodos[Algoritmos]

- **Força Bruta:** Realiza extenso número de iterações para identificar similaridades e padrões frequentes em séries ou subsequências. (Maior custo computacional)
- **Probabilísticos:** Utiliza funções de probabilidade para identificar prováveis padrões frequentes nas séries temporais ou subsequência. (Maior custo computacional)



Pesquisa em Desenvolvimento - Identificação de Motifs em Séries Espaço Temporais

- A pesquisa que está sendo desenvolvida busca estender os métodos existentes para identificação de Motifs em Séries Temporais para identificação de Motifs em Séries **Espaço Temporais**
- Estudo de Caso: Análise de dados Sísmicos



Pesquisa em Desenvolvimento - Identificação de Motifs em Séries Espaço Temporais - Formalização

Definition

A **time series** t *is a series of values* $< t_1, ..., t_m >$, *where* $|t|$ *is the number* m *of elements in* t, *and* t_m *is the most recent value in* t.

Definition

A **subsequence** r *of size* n *in a time series* t *is a series of values* $< r_1, ..., r_n >$, *such that there exist* $i_1 < i_2 < \ldots < i_n$ *integers in which* $r_1 = t_{i_1}$, $r_2 = t_{i_2}, \ldots$ $p_n = t_{i_n}$, $|r| = n$. *Formally,* $r = subseq(t, \{i_1, i_2, \ldots, i_n\})$.

When the actual set of indexes is not important during analysis, we can specify, for short, $r = subseq(t)$.

Pesquisa em Desenvolvimento - Identificação de Motifs em Séries Espaço Temporais - Formalização

Definition

A **pattern** p *is a series of values. Let* $p =< p_1 p_2 \dots p_n >$ *a pattern and* $t =< t_1 t_2 \dots t_m >$ *a time series, where* $m > n$. p *is* **included** *in* t $(p \prec t)$ *iff there exists* $r = subseq(t)$ *such that* $p \simeq r$, *i.e.,*

$p_1 \simeq r_1,\ p_2 \simeq r_2, \dots\ p_n \simeq r_n$.

Definition

Given a time series t *and support* σ, *a pattern* p *is a* **motif** *in* t, *iff* p *is included in* t *at least* σ *times. Consider a set of subsequences* R *in* t, *such that* $\forall r \in R, r = subseq(t)$. *Formally, given time series* t *and pattern* p, $motif(p, t) \leftrightarrow \exists\ R$, *such that* $\forall r \in R,\ r \simeq p \wedge |R| \geq \sigma$.

Pesquisa em Desenvolvimento - Identificação de Motifs em Séries Espaço Temporais - Formalização

A spatio-temporal series is a time series associated with 2D coordinates.

Definition

A **spatio-temporal series** s *is comprised of* x *and* y*, which stand for its coordinates and* $t =< v_1, v_2, \ldots v_m >$ *a series of values.* $s.x$ *and* $s.y$ *are the coordinates of* s*, and* $s.t$ *is the series of values of* s*.*

Pesquisa em Desenvolvimento - Identificação de Motifs em Séries Espaço Temporais - Formalização

Let D be a set of spatio-temporal series. A series of values is a motif if it is included in a sufficient number of times in D. The inclusion is defined as follows:

Definition

Given D, a set of spatio-temporal series and a support σ, a pattern p is a **motif** *if p is included in at least σ times in D.*

Pesquisa em Desenvolvimento - Identificação de Motifs em Séries Espaço Temporais - Formalização

A block is a set of series that belong to a restricted area in the 2D space. The restricted area being estimated by a given threshold.

Definition

Let $b = \{s_1, s_2, \ldots, s_k\}$ *be a set of* k *spatio-temporal series.* b *is a* **block** *iff* $\forall s_i, s_j \in b, d(s_i, s_j) < \delta$, *where* $d(s_i, s_j)$ *is any distance measure based on* $s_i.x$, $s_i.y$, $s_j.x$ *and* $s_j.y$, *and* δ *is a threshold given by the end-user.*

Pesquisa em Desenvolvimento - Identificação de Motifs em Séries Espaço Temporais - Formalização

Two blocks are contiguous if they are distinct and the minimum distance between two of their series is below a given threshold (they "touch").

Definition

Two blocks b_1 and b_2 are contiguous if $\forall s_i \in b_1, s_j \in b_2, s_i \neq s_j$ and $\exists s_1 \in b_1, s_2 \in b_2$ such that $d(s_1, s_2) \leq \tau$ where τ is the separation threshold, given by the end-user.

Pesquisa em Desenvolvimento - Identificação de Motifs em Séries Espaço Temporais - Formalização

A pattern is a tight motif if it is frequent in restricted areas only. If the area is too large, then the pattern is not a tight motif. In the meanwhile, if the area is large, but comprised of separate blocks, then the pattern remains tight.

Definition

Let tp be a pattern. tp is a **tight motif** *iff:*

1. $\exists b$, *a block such that tp is frequent in the series of b.*
2. $\forall b_i$, *a block contiguous to b, then p is not frequent in the series of* b_i.

Pesquisa em Desenvolvimento - Motifs em Séries Espaço Temporais

- Dúvidas?

Pesquisa em Desenvolvimento - Motifs em Séries Espaço Temporais

- Obrigado!
  Email: murillodutra@gmail.com