UFRRJ
UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO

Semana de Integração

CEFET/RJ

Detecção de Eventos em Séries Temporais

Eduardo Ogasawara
eogasawara@ieee.org
https://eic.cefet-rj.br/~eogasawara

Biografia

- Doutor em Engenharia de Sistemas e Computação (COPPE/UFRJ) em 2011
- Professor no EIC - CEFET/RJ
  - Departamento de Ciência da Computação
  - Curso Técnico de Informática
  - Programa de Pós-Graduação em Ciência da Computação (PPCIC)
  - Programa de Pós-Graduação em Engenharia de Produção e Sistemas (PPPRO)
- Membro do Sênior da IEEE
- Membro da SBC e ACM
- Editor Associado da IEEE Latin America Transactions

https://eic.cefet-rj.br/~eogasawara

http://lattes.cnpq.br/0528303491410251
https://www.researchgate.net/profile/Eduardo_Ogasawara
https://www.linkedin.com/in/eogasawara

Detecção de eventos

- Data
- Anomalias
- Pontos de mudança
- Motifs
- Detecção online

Predição & desvio de conceito

- Classificação
- Regressão
- Desvio de conceito

Mineração de Dados e IA Centrada em Dados

- Métodos
- Aplicações

Data Analytics Lab Team

Doutorado

Ellen Paixão
(CEFET/RJ)

Lais Baroni
(CEFET/RJ)

Lucas Giusti*
(CEFET/RJ)

Paulo Elias*
(UFF)

Mestrado

Antônio
Mello
(CEFET/RJ)

Arthur Garcia
(CEFET/RJ)

Cristiane Gea
(CEFET/RJ)

Daniel dos
Santos*
(COPPE/
UFRJ)

Edson
Sobrinho
(CEFET/RJ)

Frank Faisca
(CEFET/RJ)

Janio Lima
(CEFET/RJ)

Jéssica de
Souza
(CEFET/RJ)

Luiz Oliveira
(CEFET/RJ)

Michel Reis*
(CEFET/RJ)

Ricardo
Buçard*
(CEFET/RJ)

Thiago
Marques
(CEFET/RJ)

https://eic.cefet-rj.br/~dal/equipe/

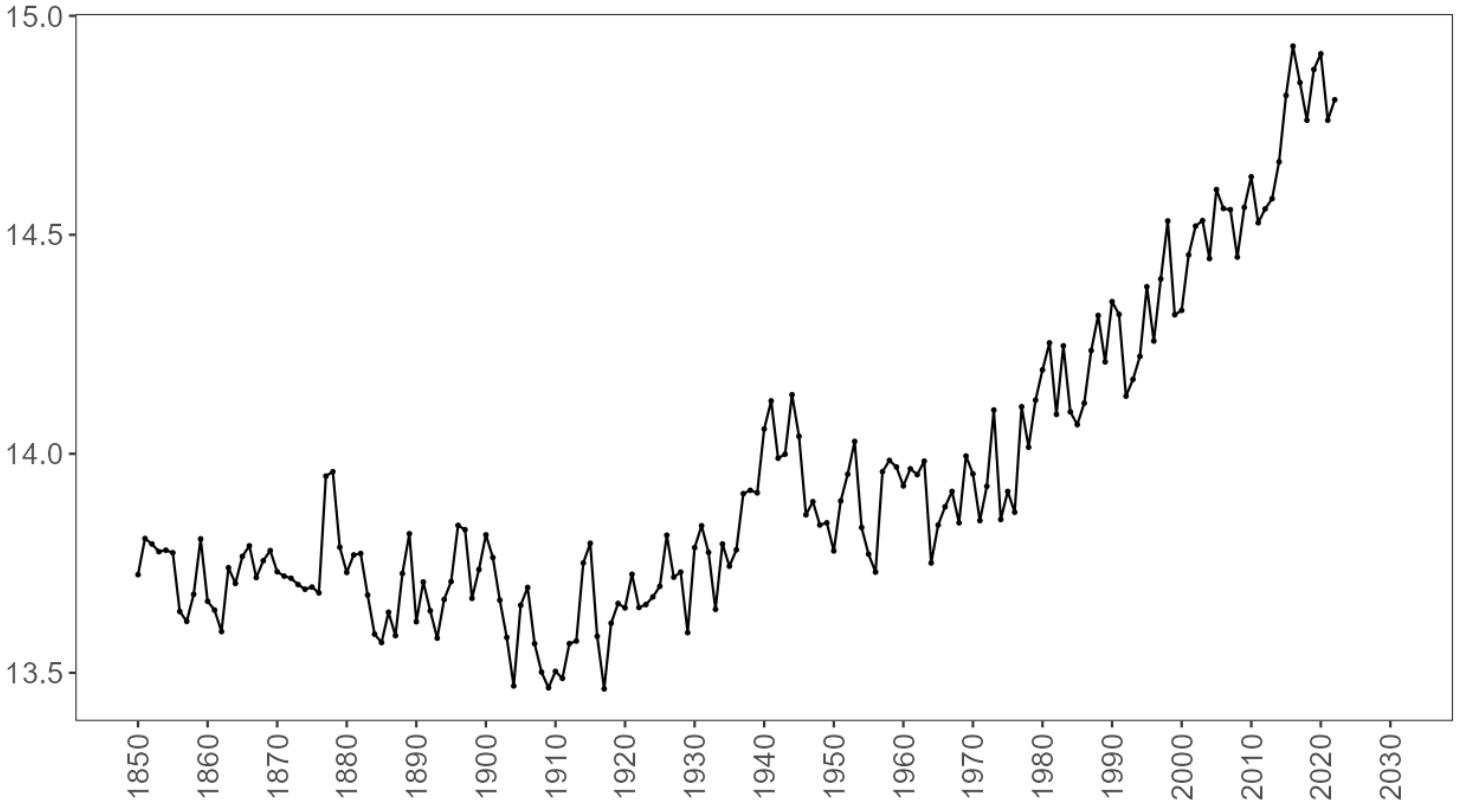
Colaboração

- CEFET/RJ
  - Docentes do PPCIC/PPPRO
    (Rafaelli Coutinho, Eduardo Bezerra, Diego Carvalho)
- Universidades e Institutos de Pesquisa
  - LNCC
    - Fabio Porto, Antônio Tadeu A. Gomes
  - Fiocruz
    - Cristiano Boccolini, Marcel Pedroso
  - COPPE/UFRJ
    - Marta Mattoso, Geraldo Zimbrão, Geraldo Xexéo
  - UFF
    - Daniel Oliveira, Leonardo Murta, Vanessa Braganholo
- Internacionais
  - INRIA / University of Montpellier
    - Patrick Valduriez, Esther Pacitti, Florent Masseglia
  - University of Nottingham
    - Jonathan Garibaldi, Chao Chen
  - Università di Camerino
    - Carlo Lucheroni

https://www.inria.fr/en/zenith
http://dexl.lncc.br

Eventos

- Eventos estão relacionados a fenômenos observados nas séries temporais
- Um ponto ou um intervalo onde ocorre uma mudança significativa no comportamento da série temporal
- Normalmente tem algum significado no domínio



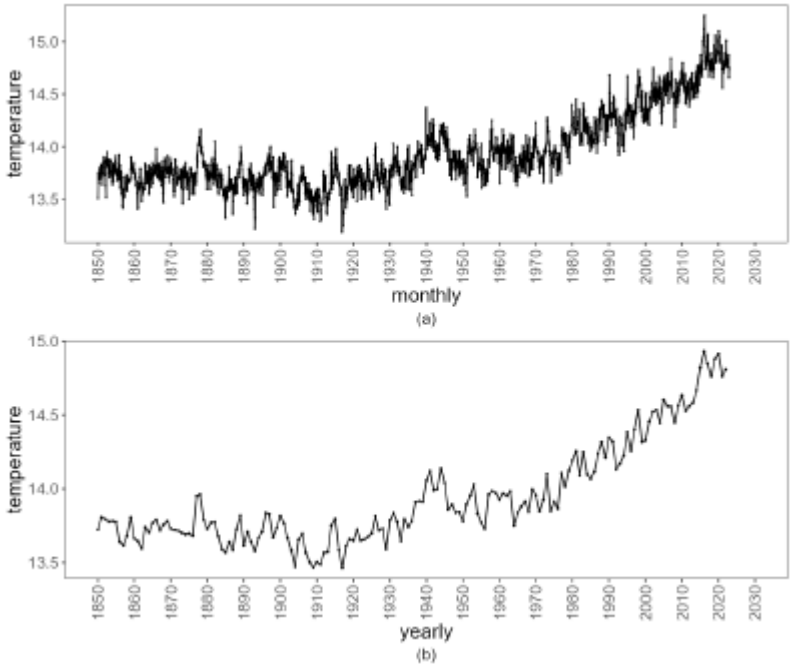
Detecção de Eventos

- Processo de identificação destes eventos
- Importante para monitoração e vigilância
  - Industria, sísmica, exploração e petróleo, epidemiologia, clima
- Há muitos estudos, mas ...
  - Focados em tipos específicos de eventos
  - Falta uma visão holística do problema

Básico de análise de series temporais

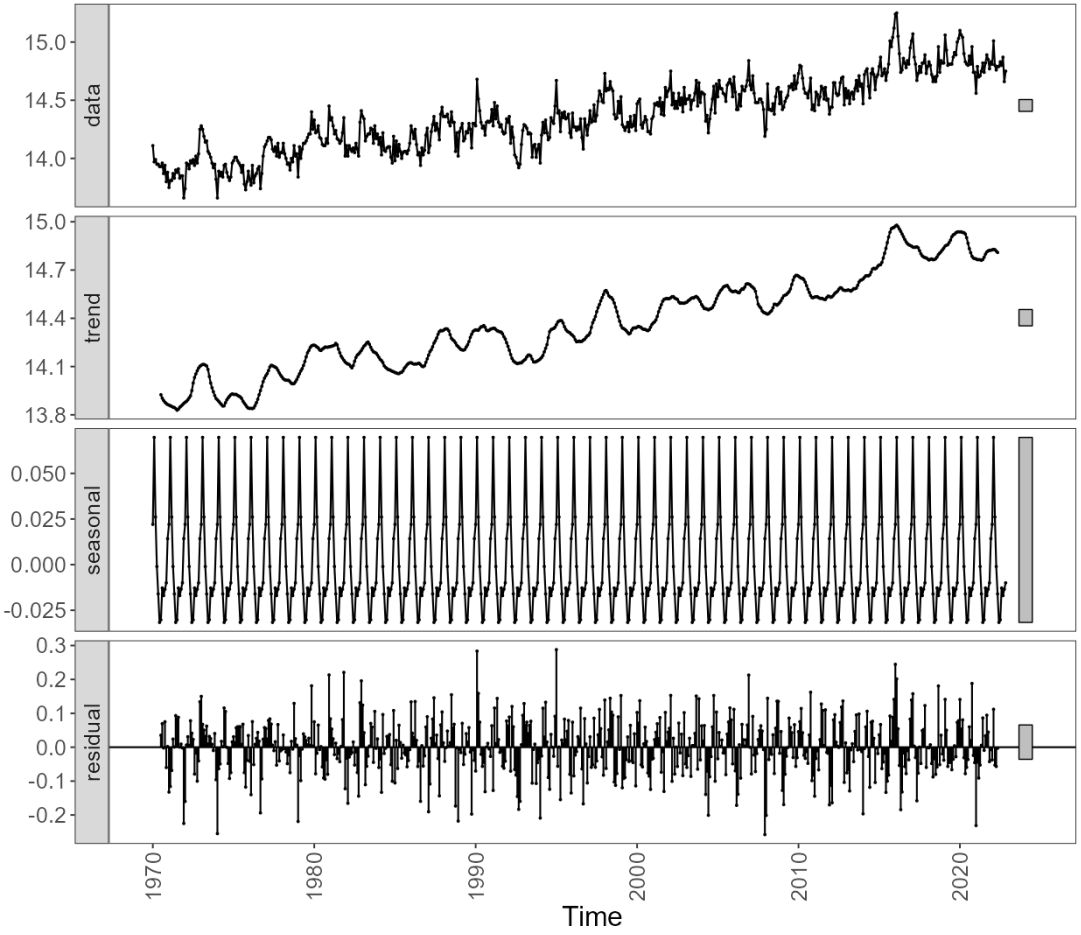
Séries temporais

- São sequências de observações
  - $\mathrm{X} = < x_1, x_1, \ldots, x_n >$
- Univariadas ou multivariadas
- Frequência (regularidade de coleta)
  - Séries diárias, semanais, mensais



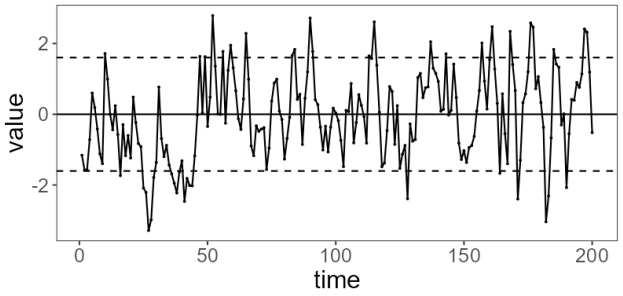
Decomposição de séries temporais

- Série temporal pode ser decomposta em tendência, sazonalidade e ruído
- $x_t = \beta_t + \pi_t + \omega_t$

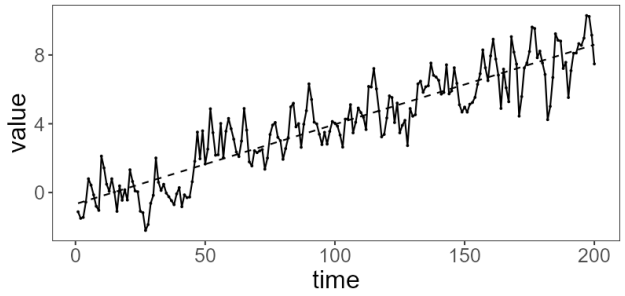


Estacionariedade

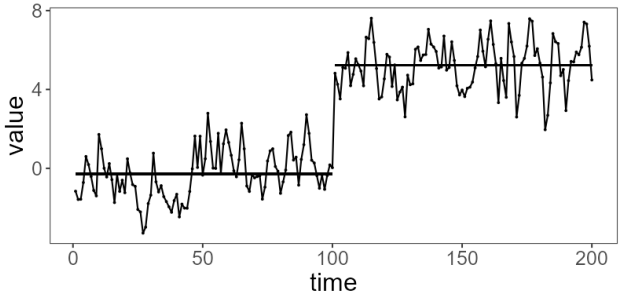
- Média constante independente do tempo
- Variância constante independente do tempo
- Autocovariância $\gamma(X_s, X_t)$ só depende de $|s - t|$



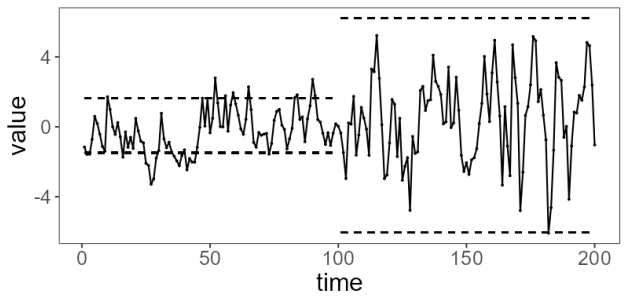
(a) stationary



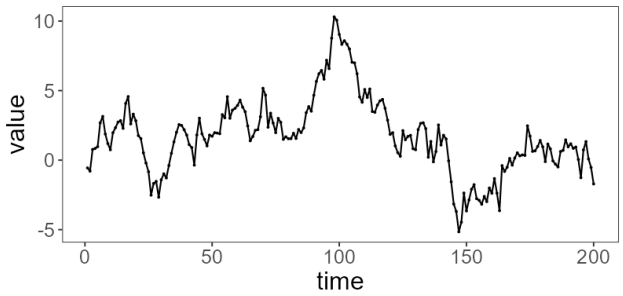
(b) trend stationary



(c) level stationary



(d) heteroscedastic

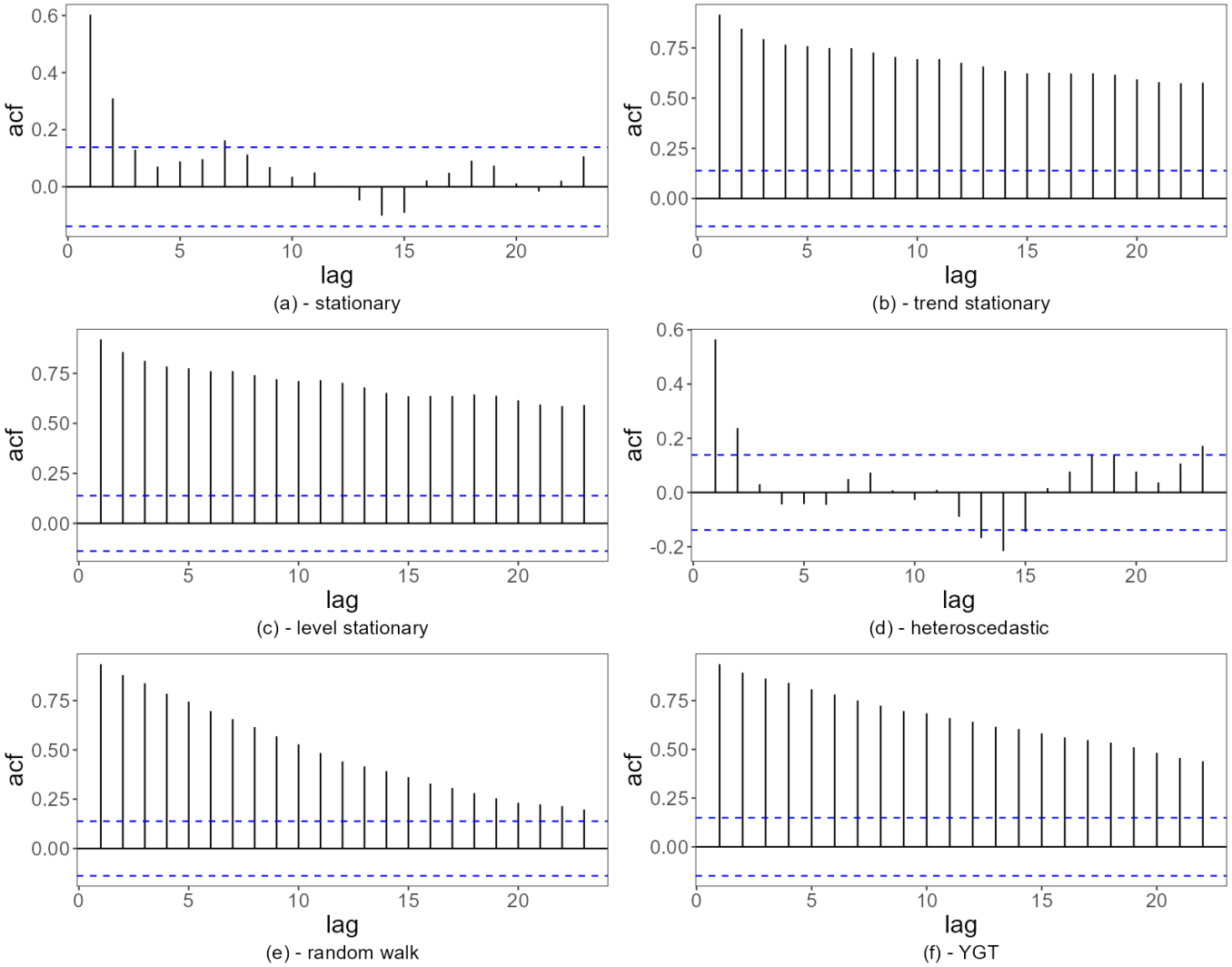


(e) random walk

Time series	ADFT	PPT	BPT
stationary			
trend stationary			
level stationary	X		
heteroscedastic			X
difference stationary	X	X	X
YGT			X

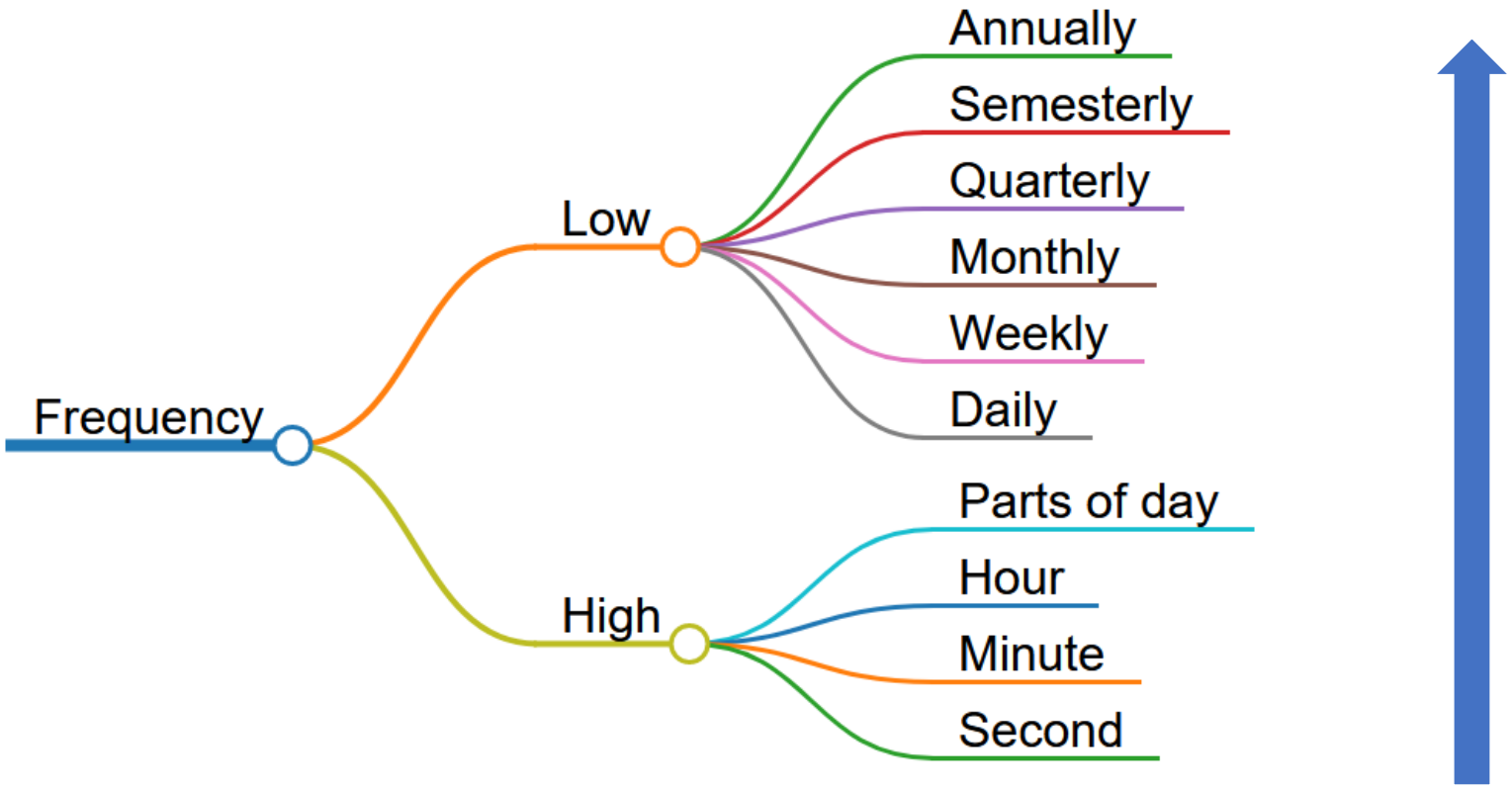
Autocorrelação

- Mede a relação entre as observações e seus termos defasados



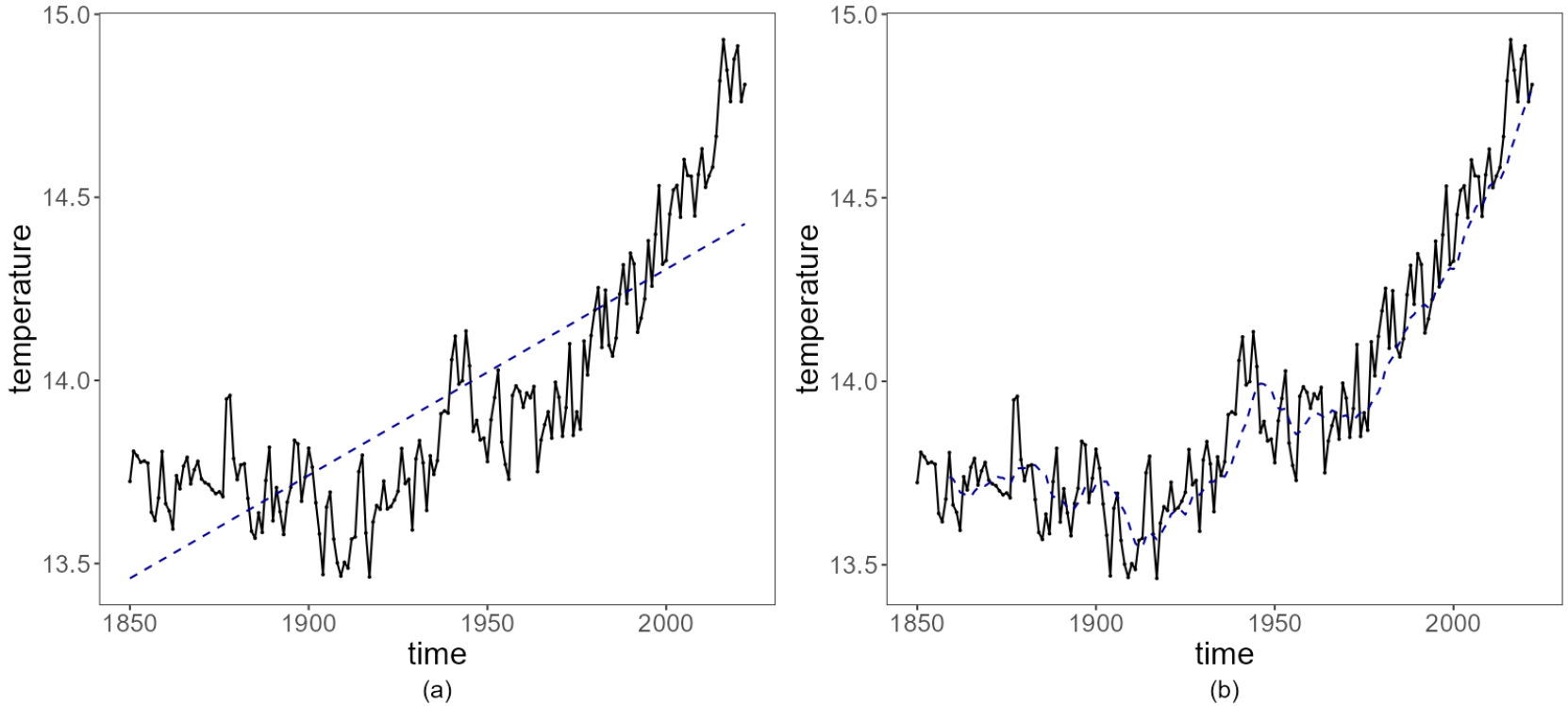
Agregação temporal

- Transformação de série de maior frequência para menor frequência



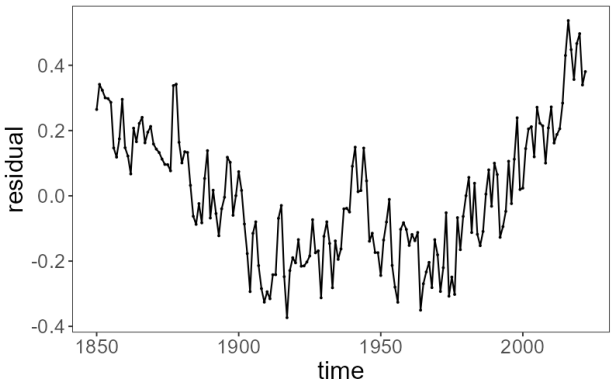
Componente de tendência

- Regressão linear
- Média móvel

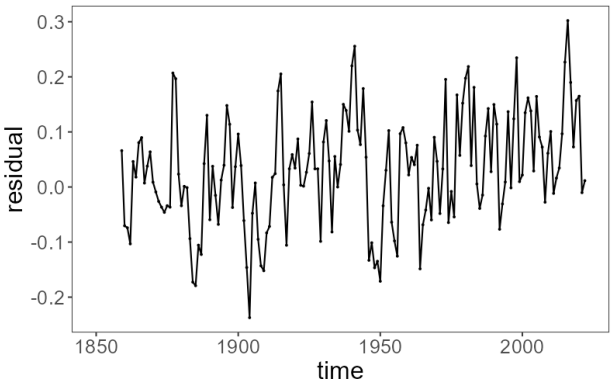


Tratamento de tendência

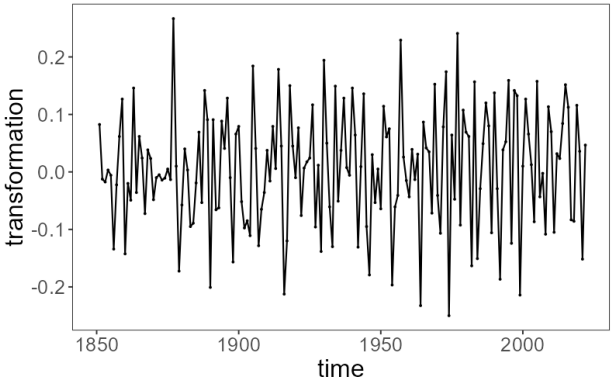
- Resíduo de regressão linear
- Resíduo de média móvel
- Diferenciação
- Variação percentual



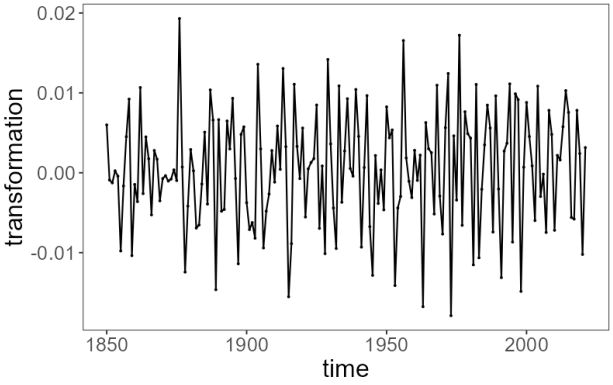
(a) LR trend removal



(b) MAS trend removal



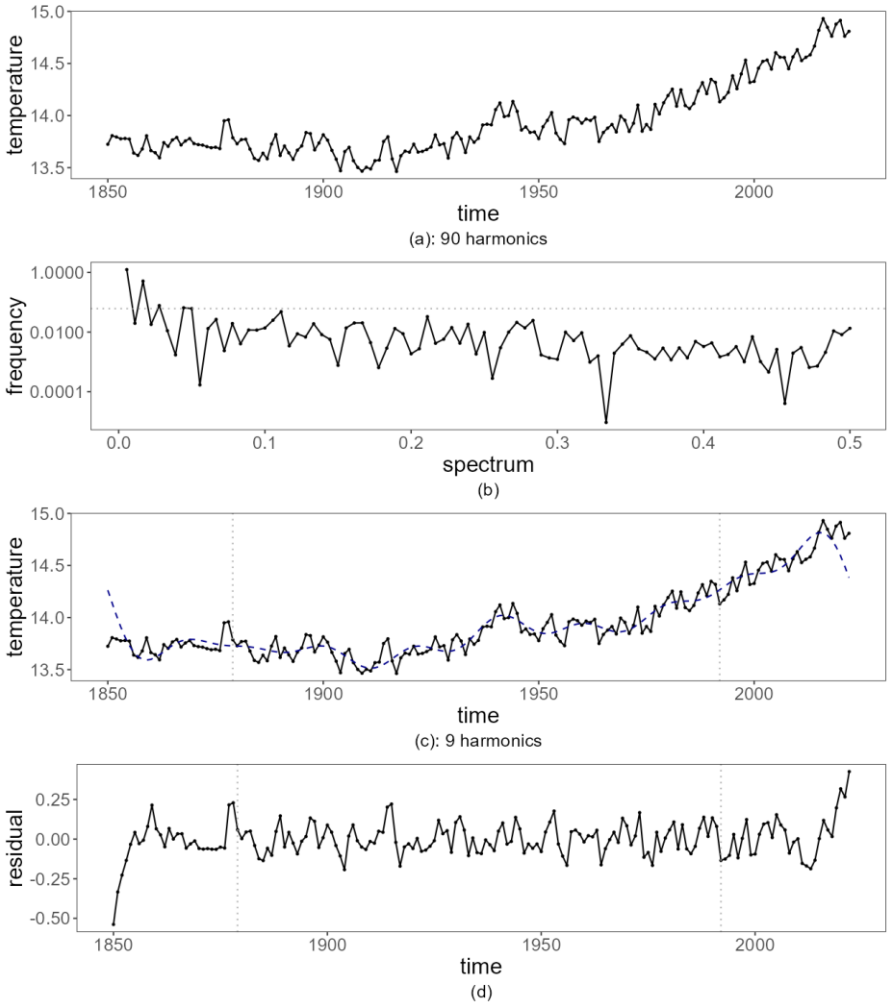
(c) First order differencing



(d) PCT

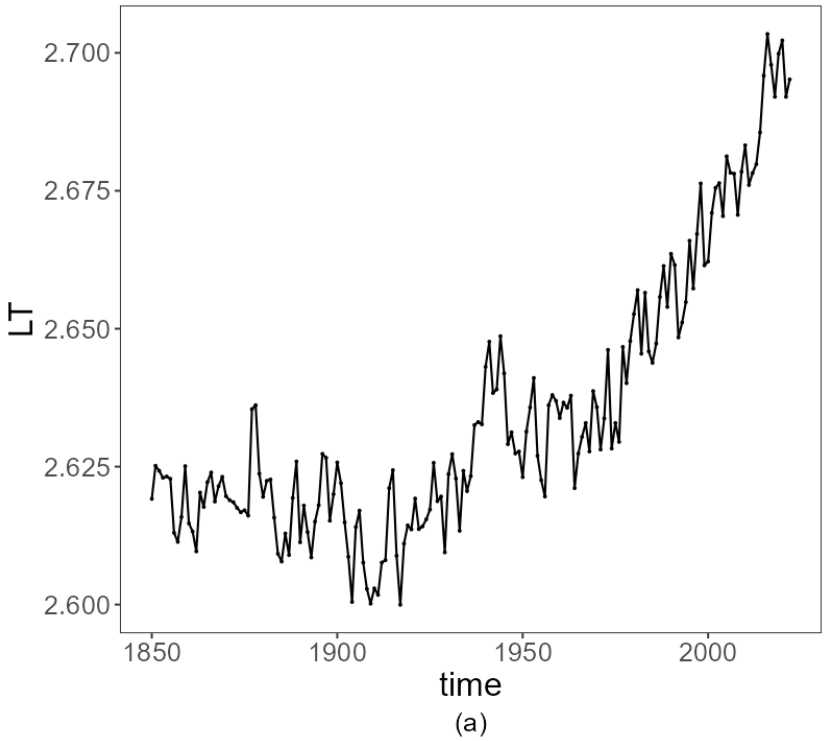
Decomposição de tendência no domínio da frequência

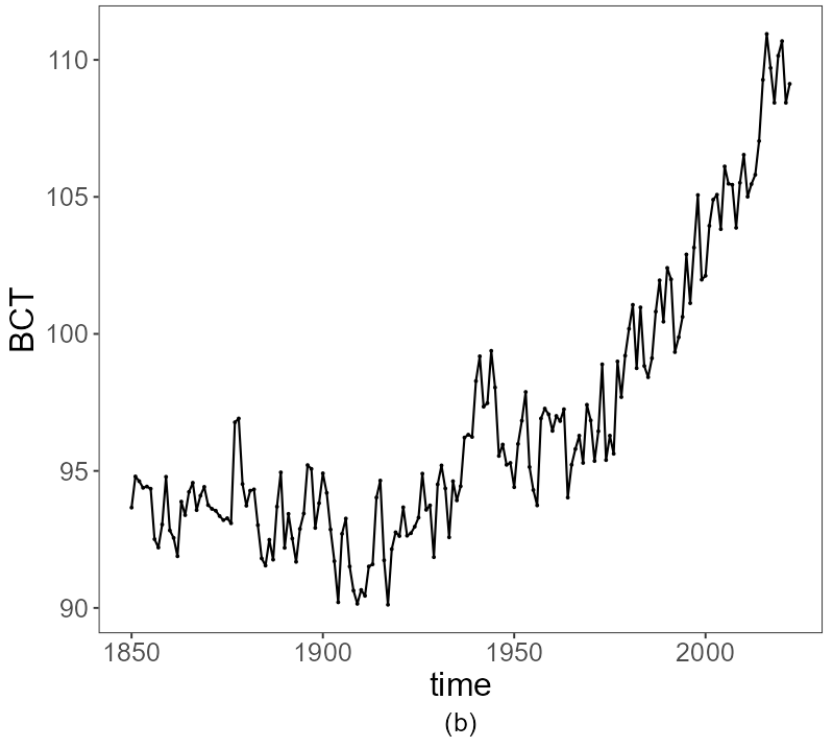
- Transformada rápida de Fourier
- Wavelet
- Empirical Mode Decomposition



Tratamento de variância

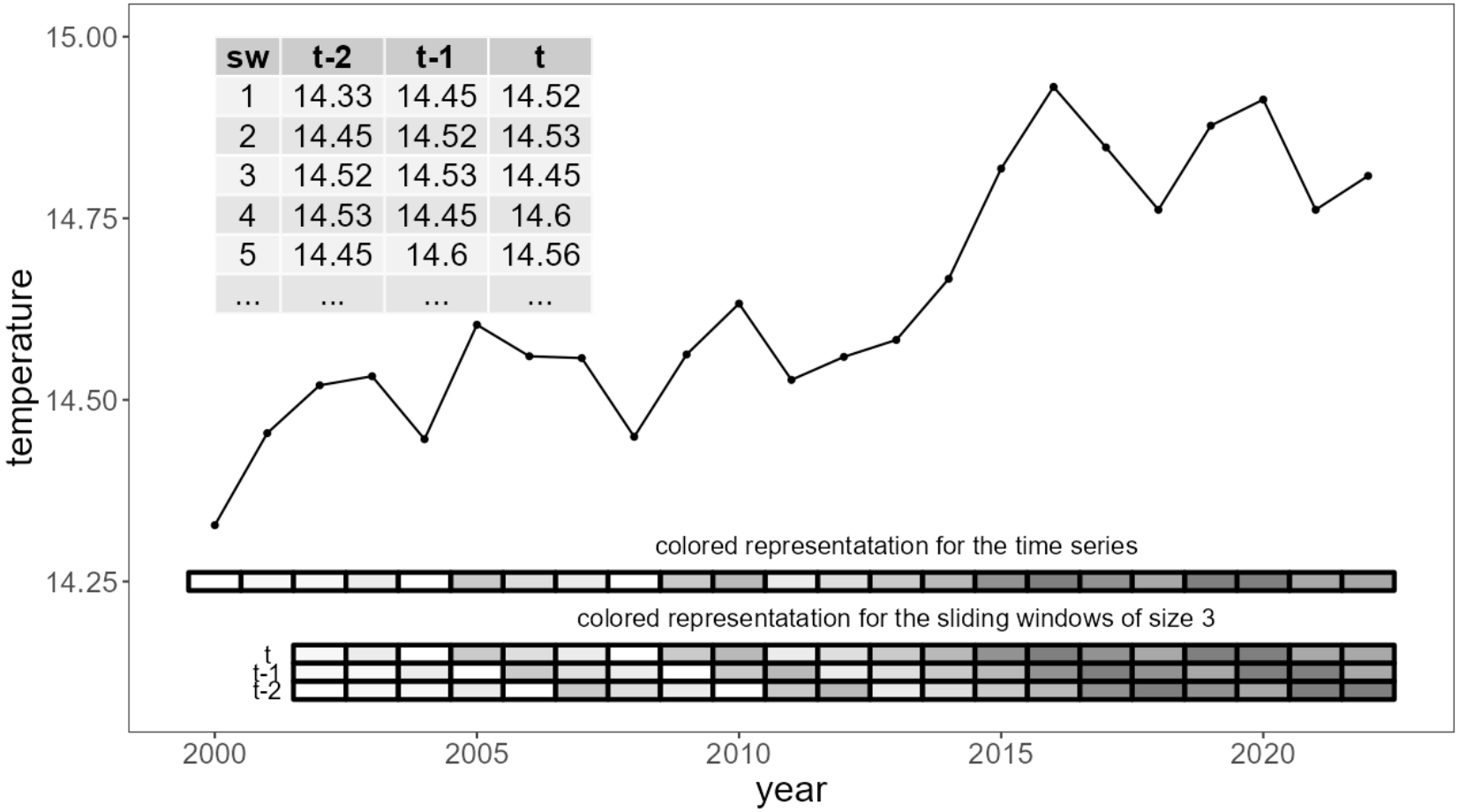
- Log (natural)
- BCT $\hat{x}_t = \begin{cases} (x_t^{\lambda} - 1)/\lambda, & \lambda \neq 0, \\ \log x_t, & \lambda = 0. \end{cases}$





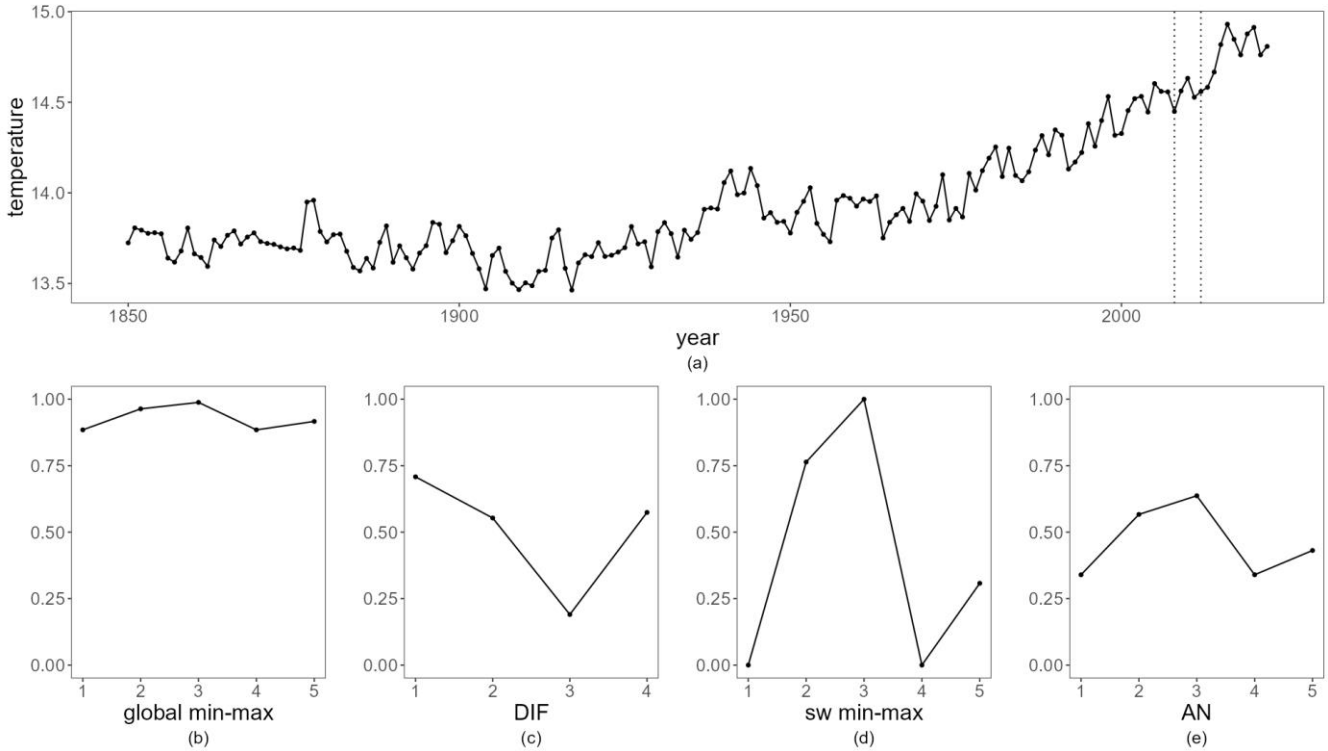
Janelas deslizantes

- Processo de analisar a série temporal a partir de subsequências
- Base para análise via aprendizado de máquina



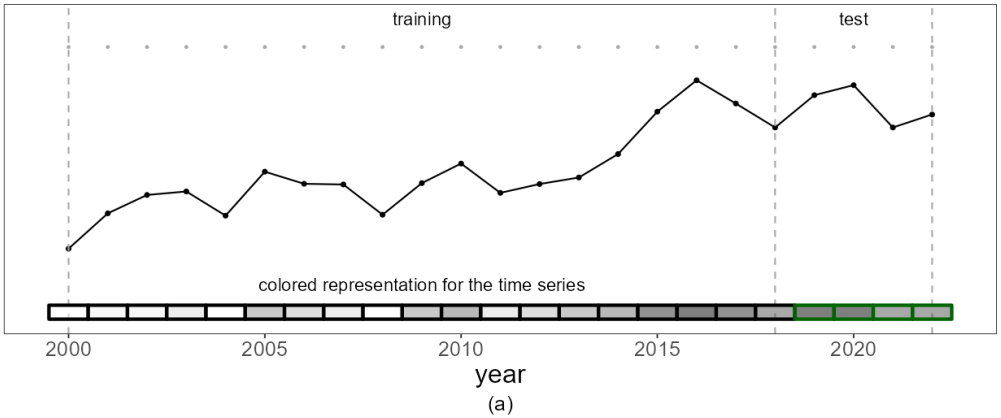
Normalização

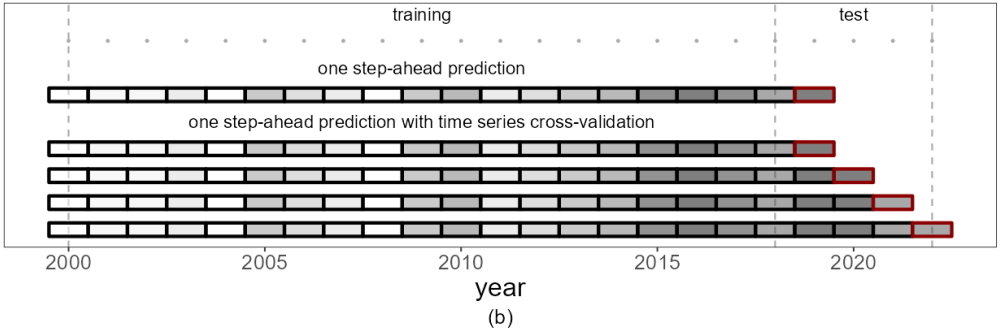
- Global Min-Max: $y_t = \frac{x_t - x_{min}}{x_{max} - x_{min}}$
- Z-Score : $y_t = \frac{x_t - \bar{X}}{\sigma_X}$
- SW - Min-Max
- Normalização adaptativa

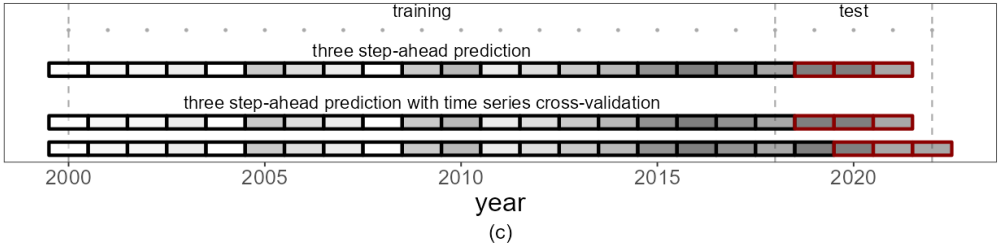


Predição de series temporais

- Separação de treino e teste
- Validação cruzada em séries temporais
- Predição passos à frente

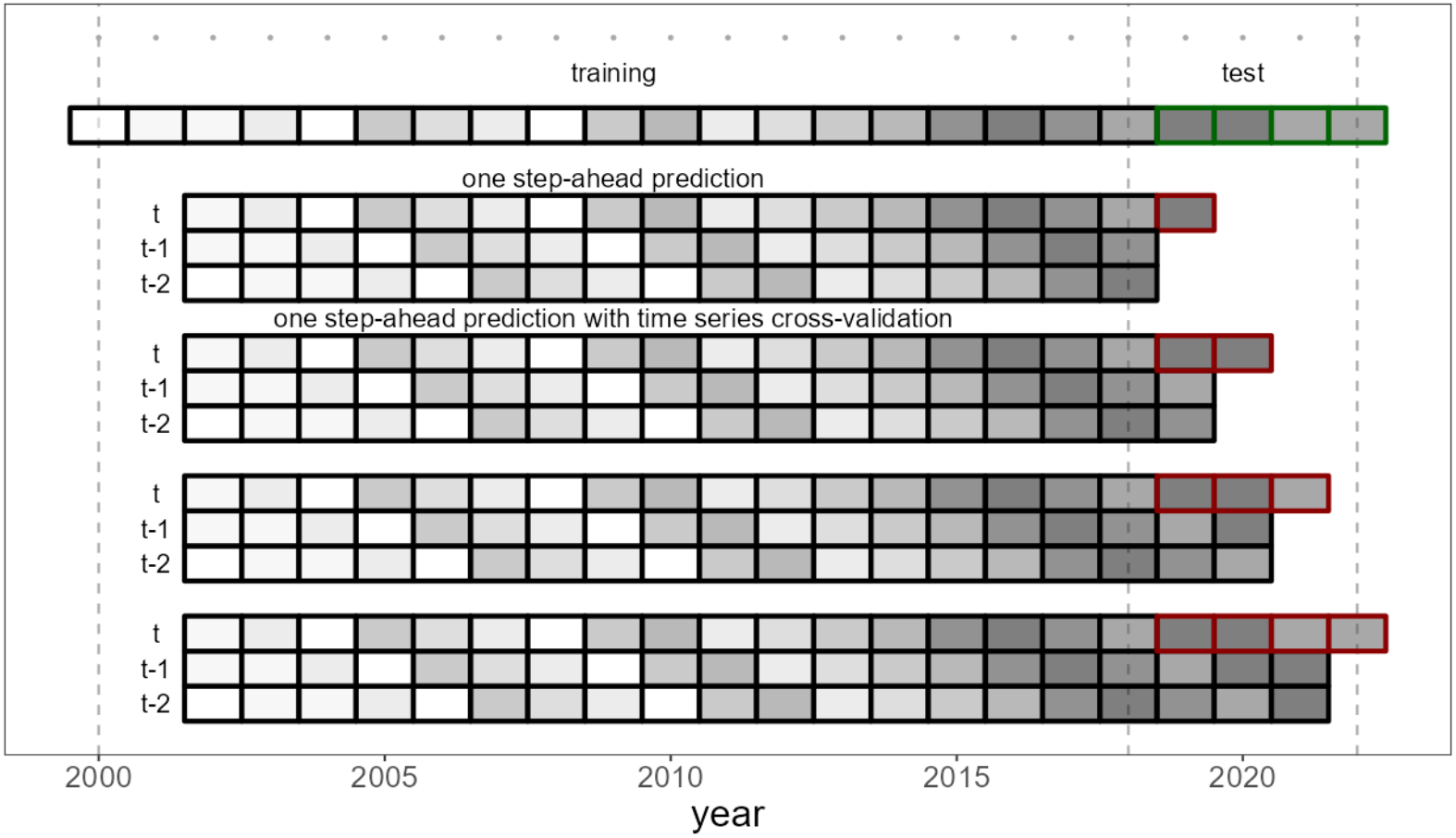




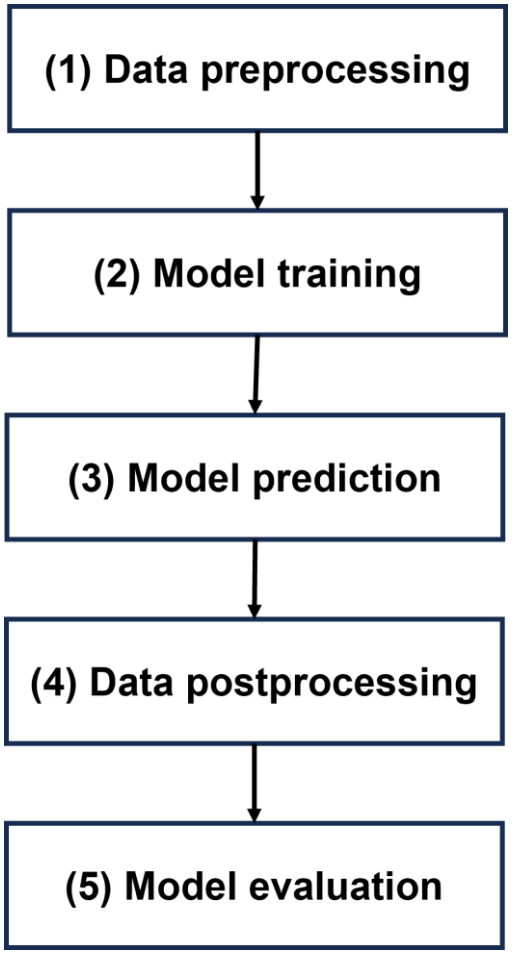


Predição de series temporais usando janelas deslizantes

- Separação de treino e teste
- Validação cruzada em séries temporais

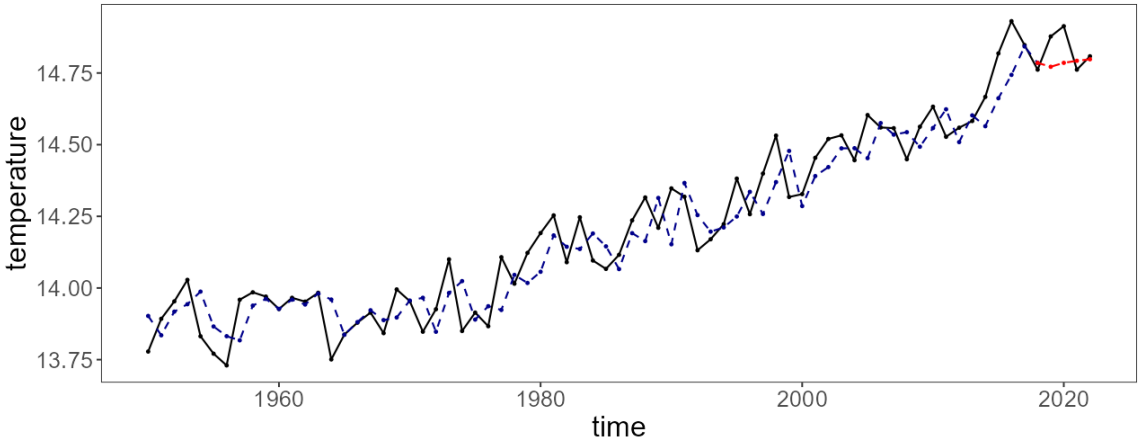


Processo de predição

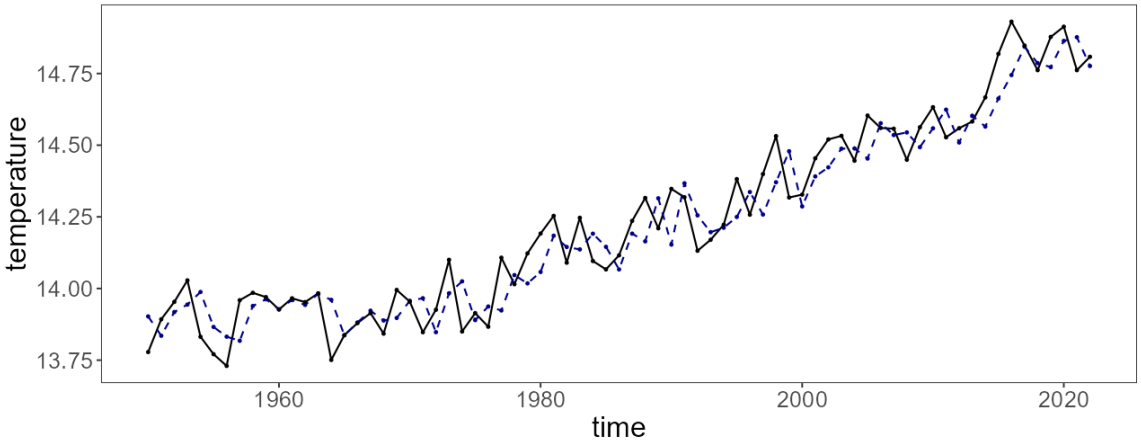


ARIMA

- ARIMA(p, d, q)

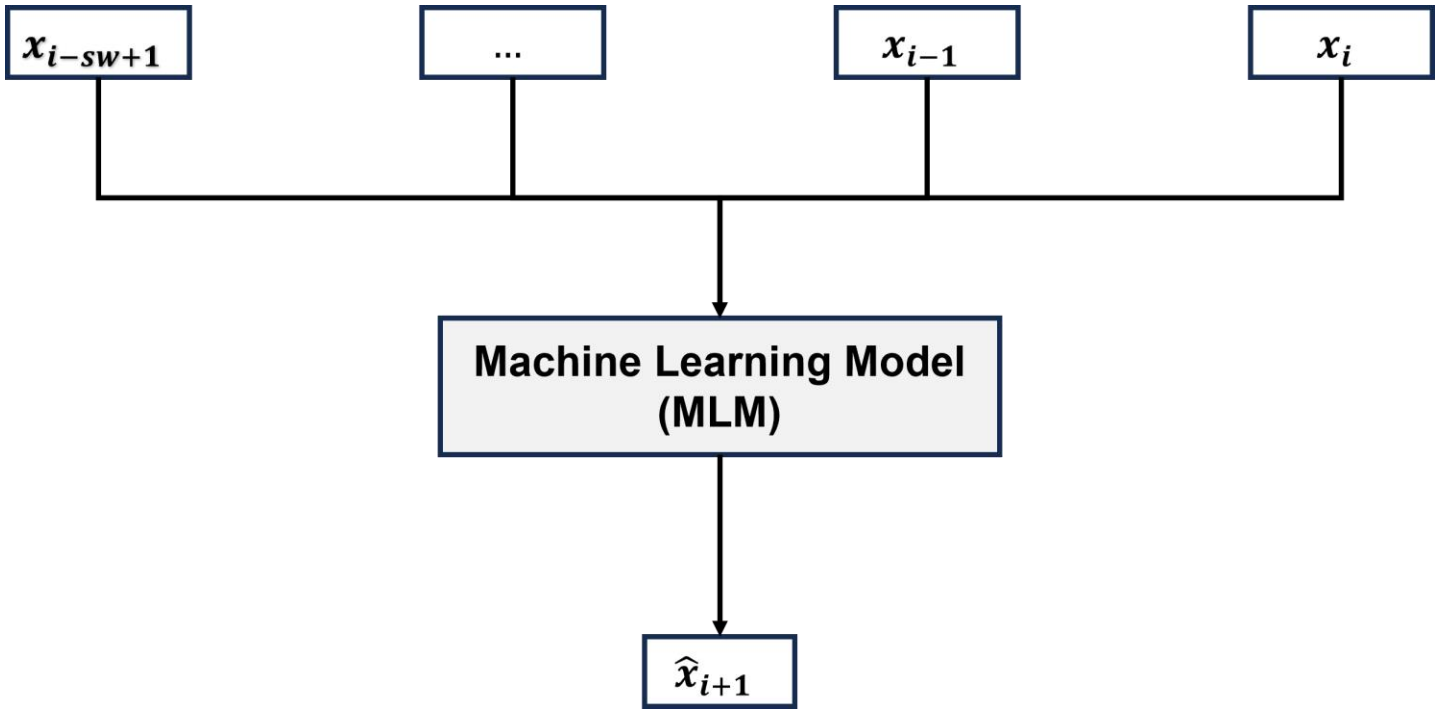


(a) ARIMA(1, 1, 3) four-step-ahead prediction



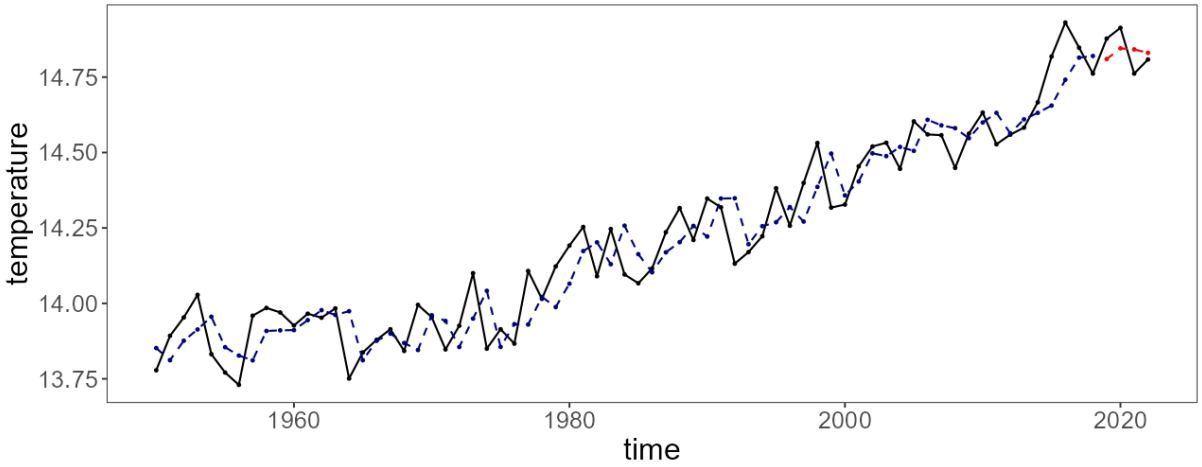
(b) ARIMA(1, 1, 3) model adjustment

Predição com aprendizado de máquina

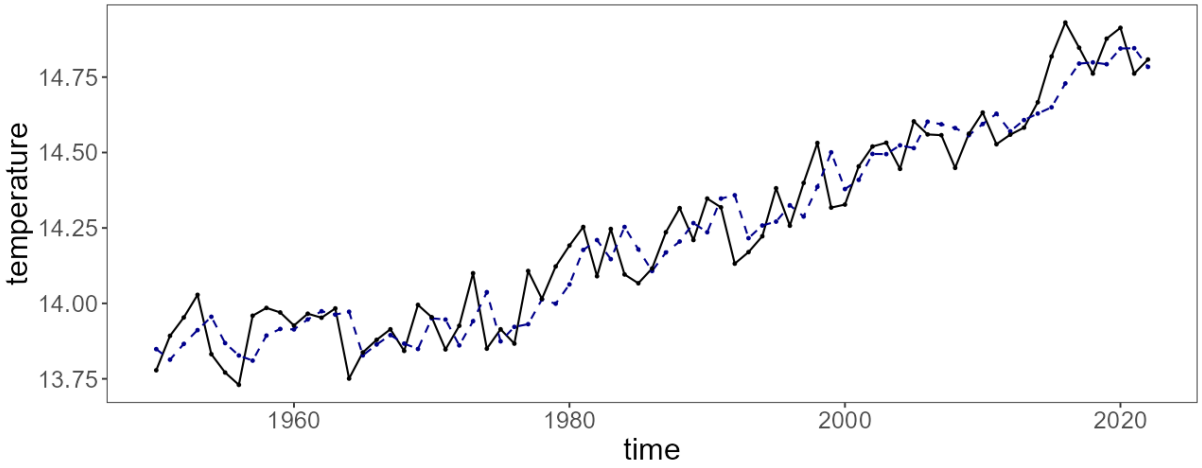


Aprendizado de máquina

- EML, MLP, RRF, SVR, Conv1D, LSTM



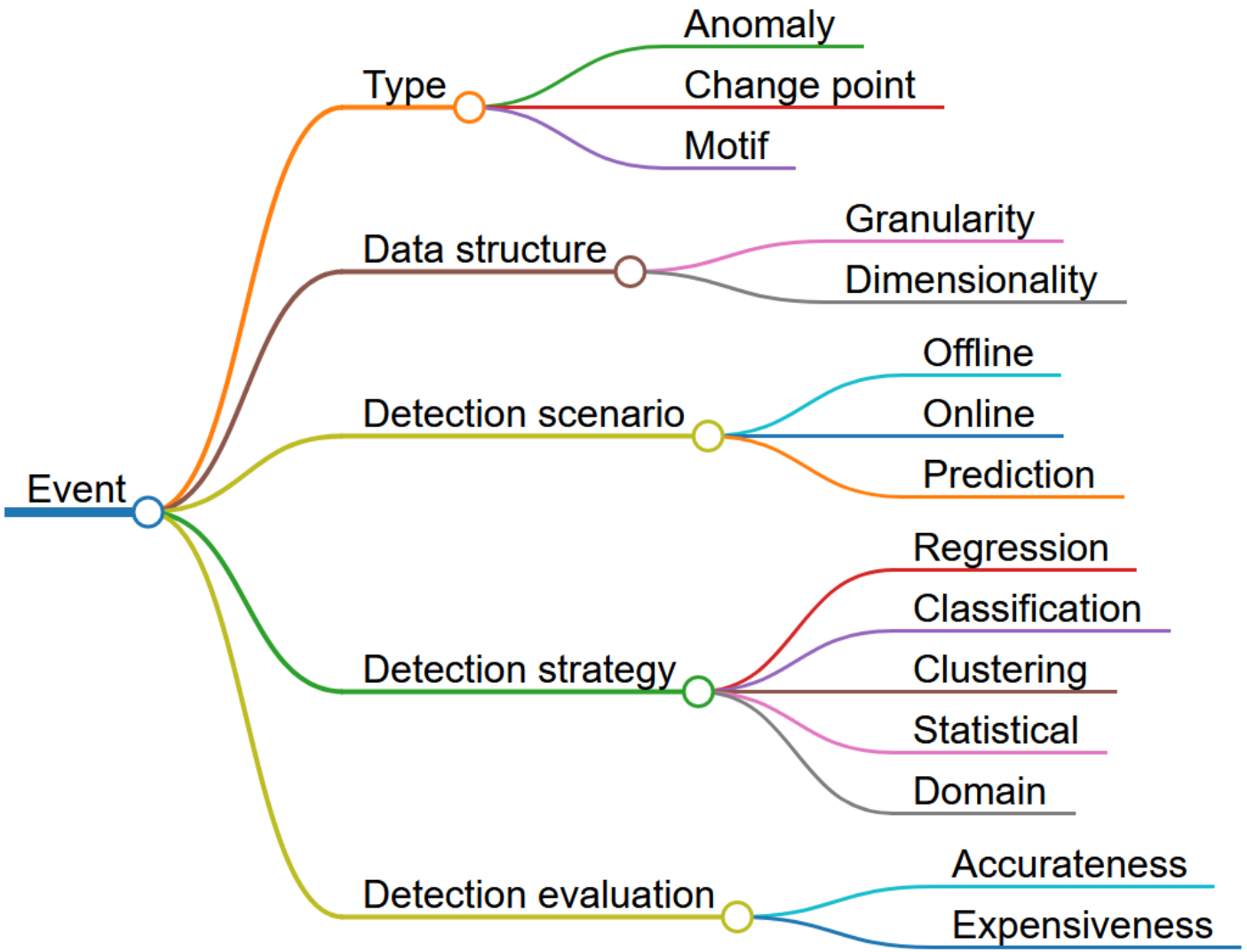
(a) LSTM four-step-ahead prediction



(b) LSTM model adjustment

Detecção de eventos

Detecção de eventos

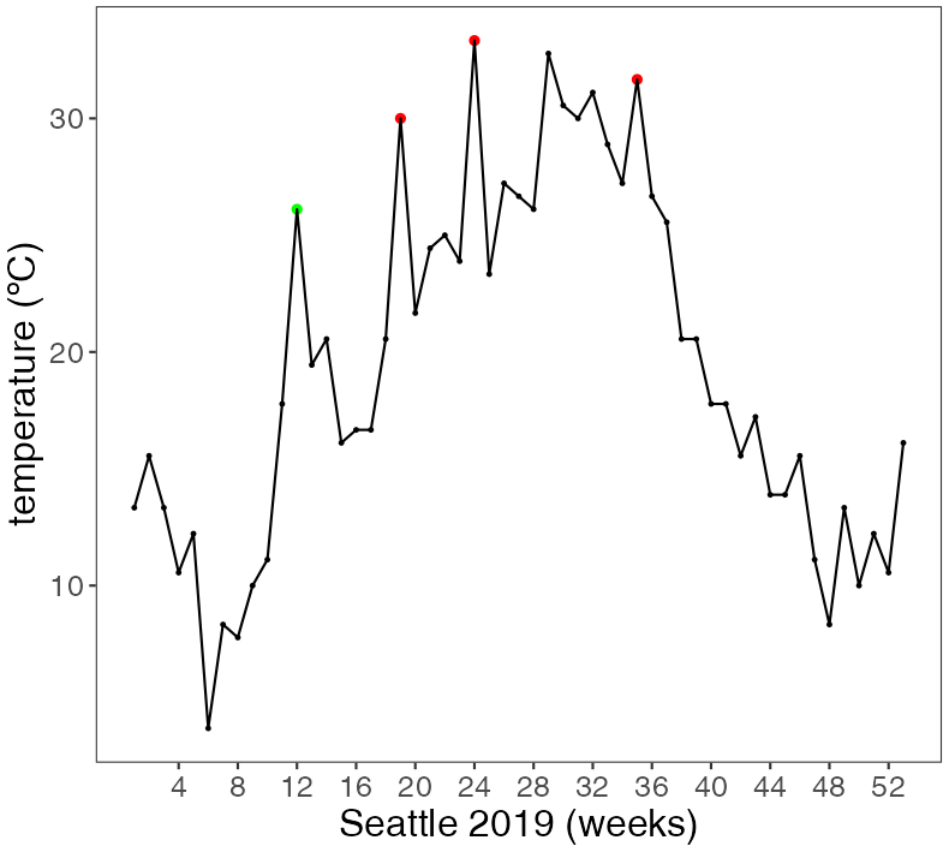


Evento

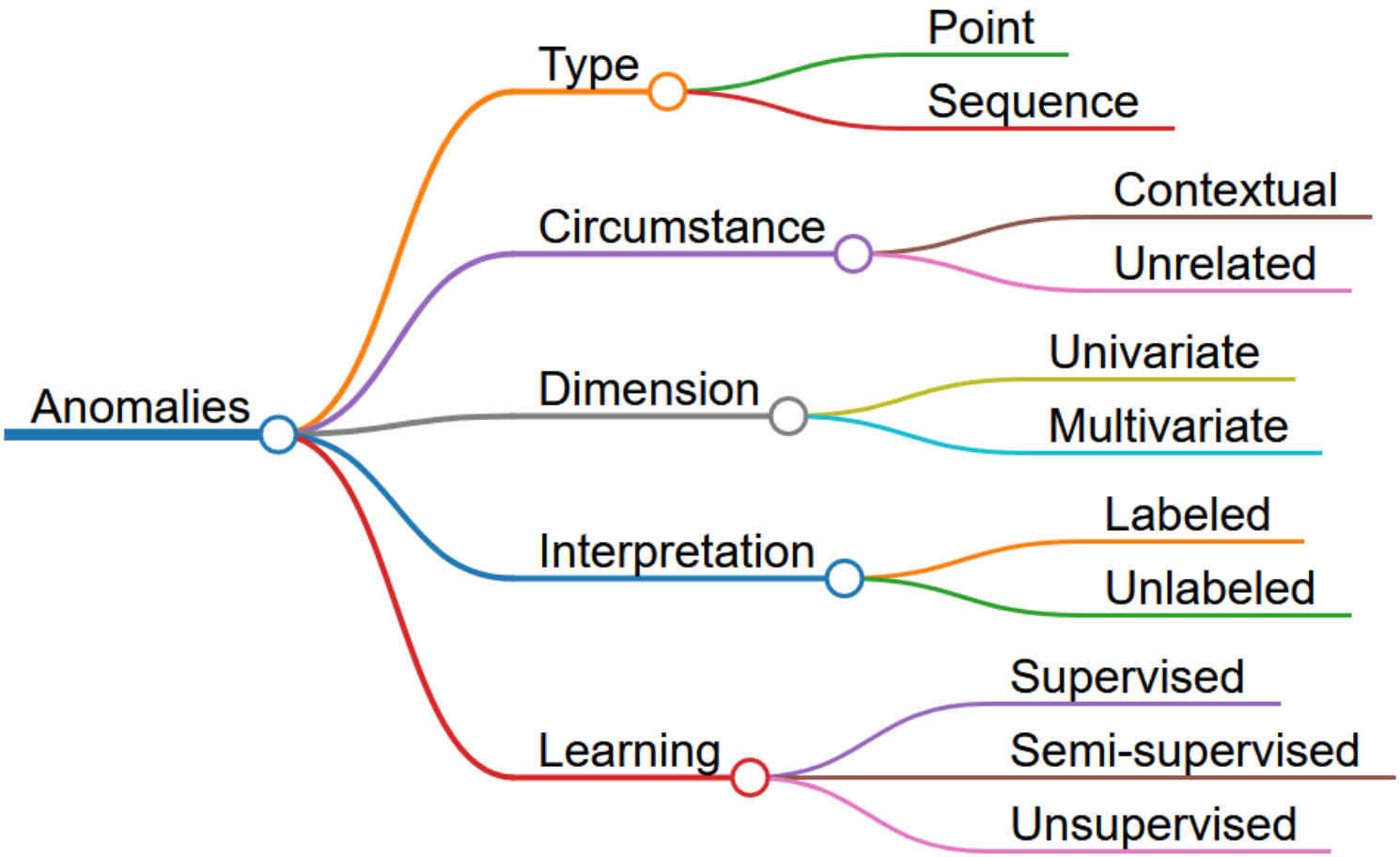
- Componente de série temporal
- $tc(x_t) = \begin{cases} tc_o(x_t) = x_t \\ tc_{tr}(x_t) = \overline{x_t} \\ tc_v(x_t) = \sigma x_t \end{cases}$
- Valor esperado de termos autoregressivos prévios
  - $ep(x_t, k) = E(tc(x_t) \mid tc(x_{t-k}), \ldots, tc(x_{t-1}))$
- Valor esperado de termos autoregressivos posteriores
  - $ef(x_t, k) = E(tc(x_t) \mid tc(x_{t+1}), \ldots, tc(x_{t+k}))$
- Evento
  - $e(X, k, \sigma) = \{t, \begin{array}{r} |tc(x_t) - ep(x_t, k)| > \sigma \\ \vee\ |tc(x_t) - ef(x_t, k)| > \sigma \\ \vee\ |ep(x_t, k) - ef(x_t, k)| > \sigma \end{array}$

Anomalias

- Um padrão ou observação que não está de acordo com o comportamento esperado [1]
- Pode ser categorizado como pontual, contextual ou coletivo
- $a(X, k, \sigma) = \{t, |tc(x_t) - ep(x_t, k)| > \sigma \wedge |tc(x_t) - ef(x_t, k)| > \sigma$

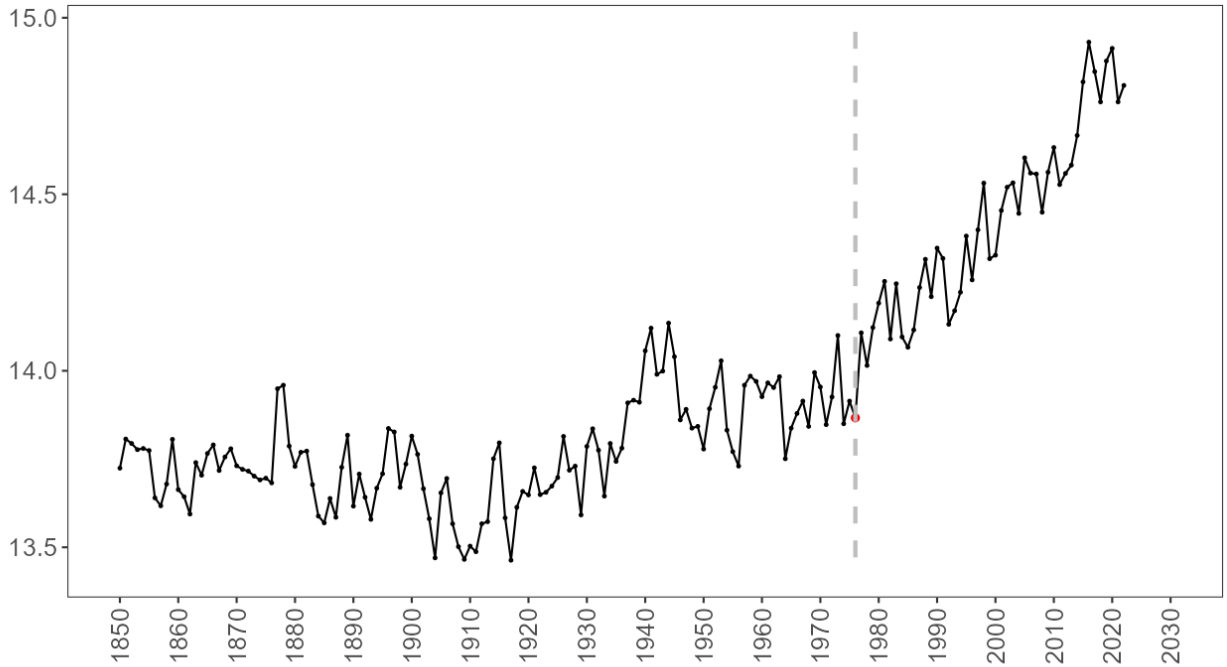


Anomalia

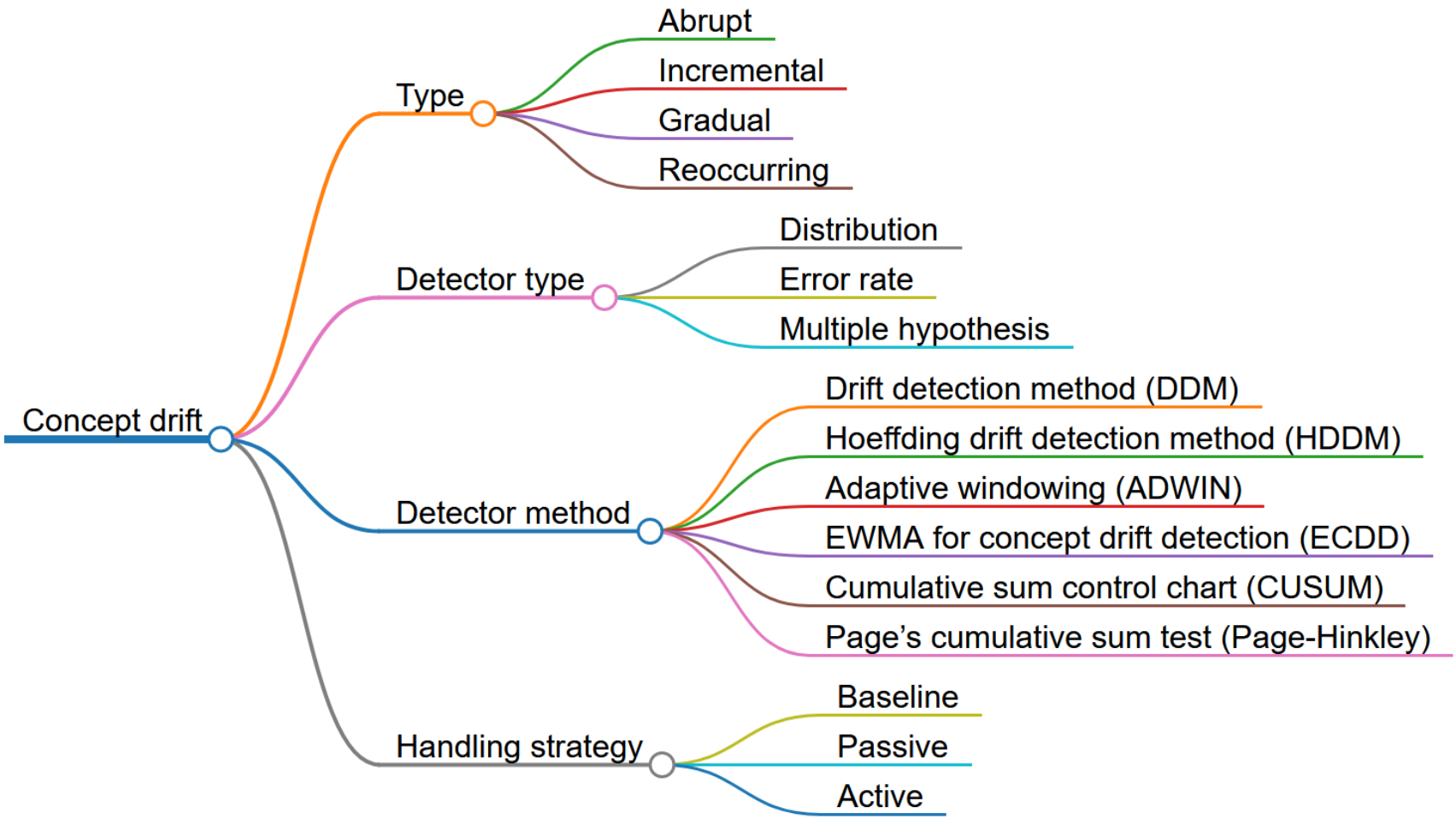

Anomalies
Type
Point
Sequence
Circumstance
Contextual
Unrelated
Dimension
Univariate
Multivariate
Interpretation
Labeled
Unlabeled
Learning
Supervised
Semi-supervised
Unsupervised

Pontos de mudança

- Pontos (ou intervalos de tempo) que indicam mudanças significativas no comportamento da série temporal [1]
- Eles separam diferentes estados no processo que gera a série temporal
- $cp(X, k, \sigma) = \{t, \begin{matrix} |tc(x_t) - ep(x_t, k)| > \sigma \veebar |tc(x_t) - ef(x_t, k)| > \sigma \\ \vee\, |ep(x_t, k) - ef(x_t, k)| > \sigma \end{matrix}$



Ponto de mudança



Categorização de desvio de conceito

- Abruto (a), Incremental (b), Gradual (c), Recorrente (d)
- Todos são exemplos de não-estacionariedade

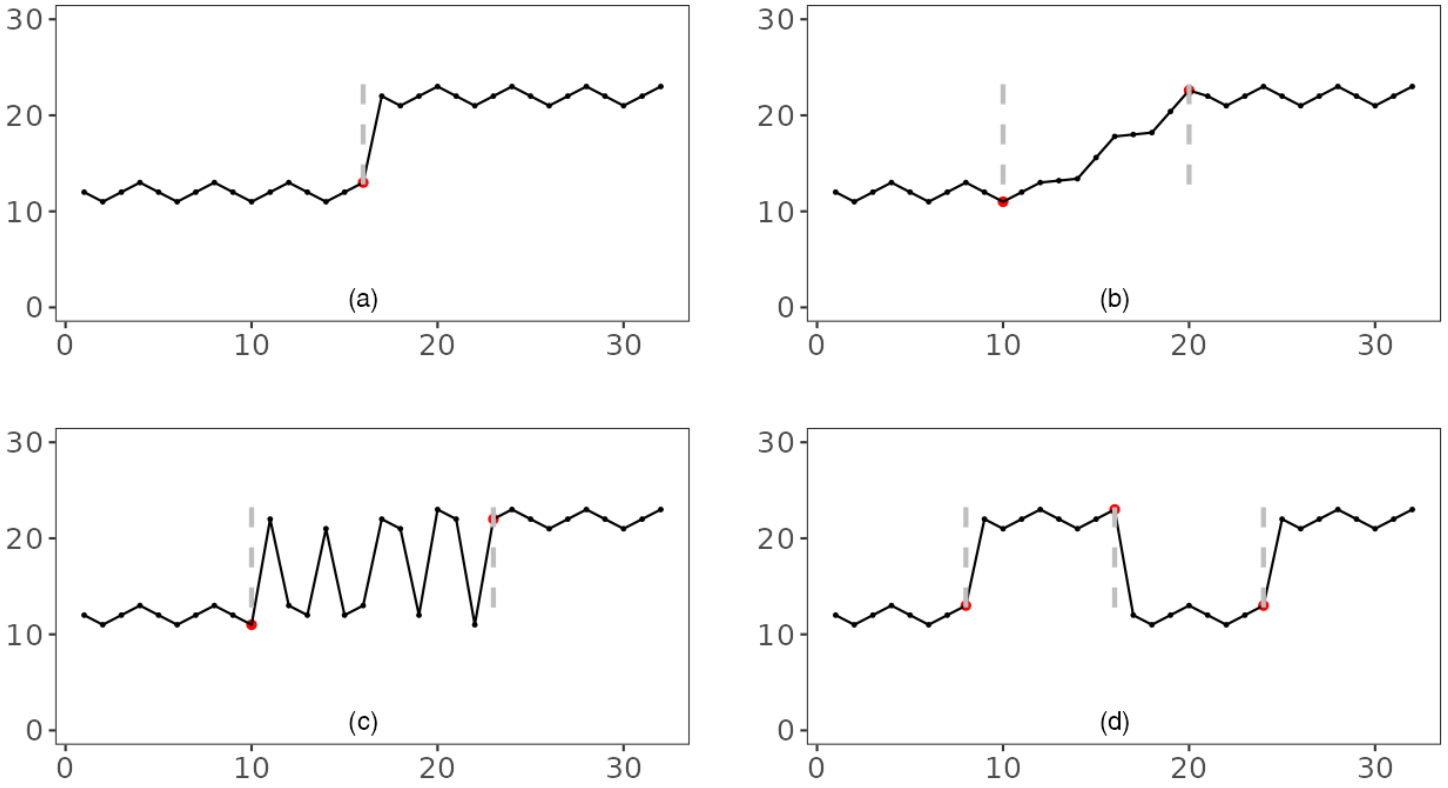
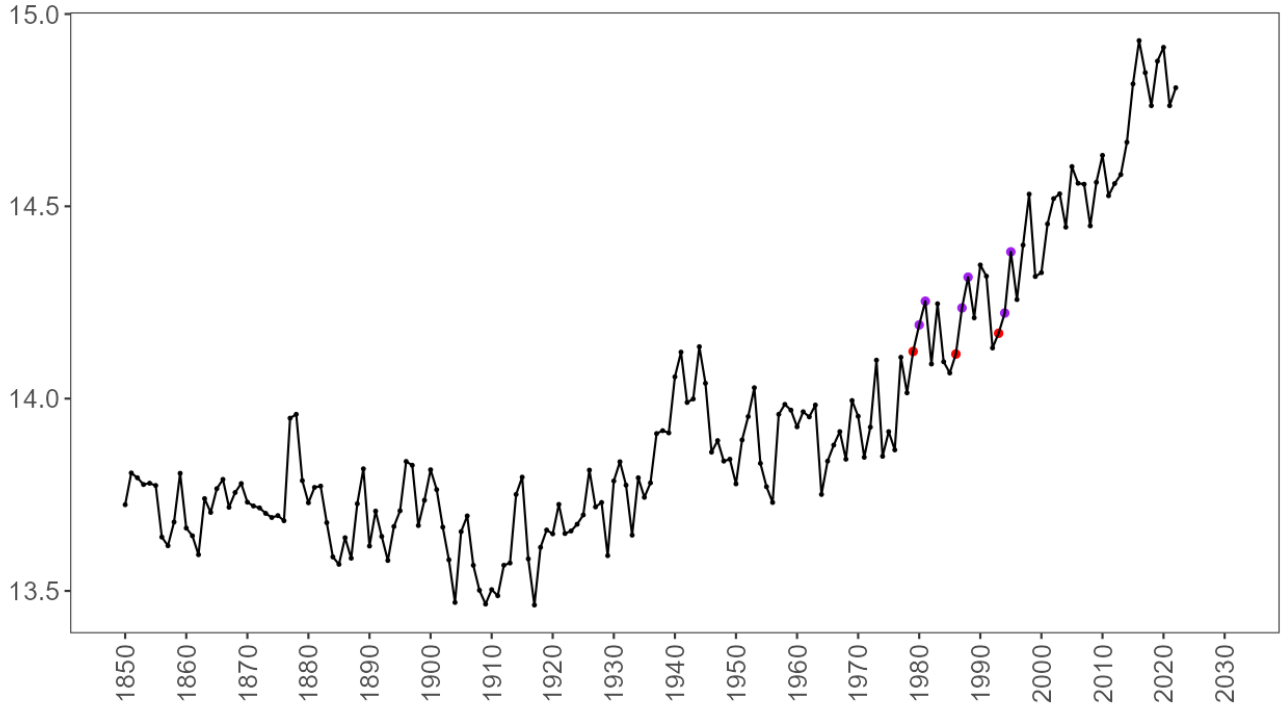


Image source: https://towardsdatascience.com/real-time-time-series-anomaly-detection-981cf1e1ca13

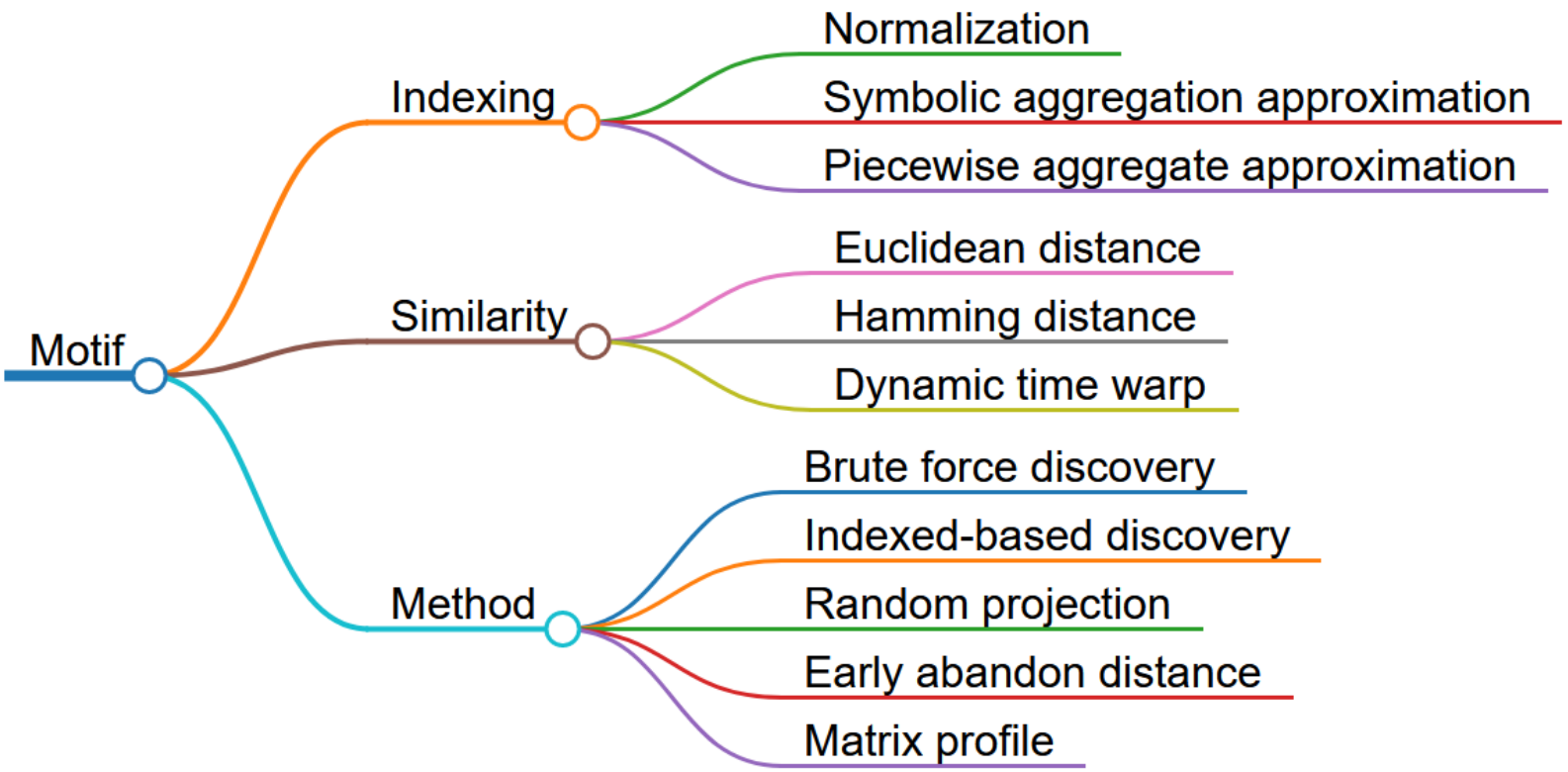
[1] G. Ditzler, M. Roveri, C. Alippi, e R. Polikar, 2015, Learning in Nonstationary Environments: A Survey, *IEEE Computational Intelligence Magazine*, v. 10, n. 4, p. 12–25.
[2] https://youtu.be/woRmeGvOaz4

Motifs

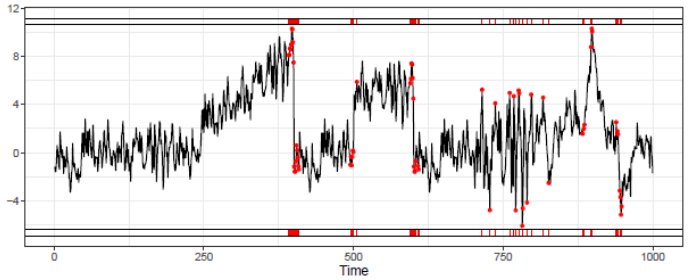
- Um padrão (desconhecido) que ocorre um número significativo de vezes em séries temporais [1, 2, 3]



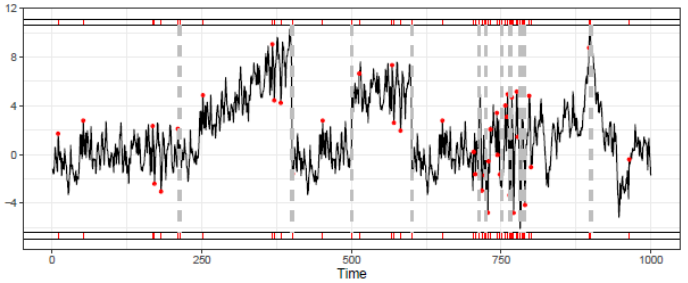
Motifs


Motif
Indexing
Normalization
Symbolic aggregation approximation
Piecewise aggregate approximation
Similarity
Euclidean distance
Hamming distance
Dynamic time warp
Method
Brute force discovery
Indexed-based discovery
Random projection
Early abandon distance
Matrix profile

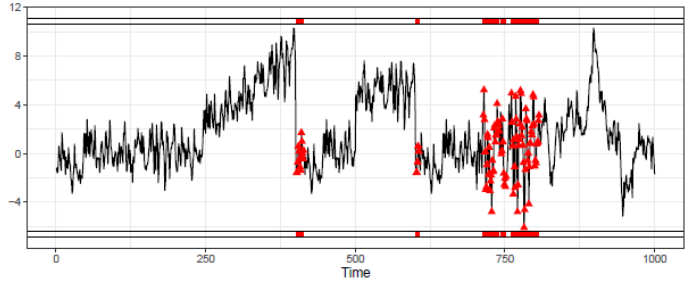
As múltiplas faces da detecção de eventos



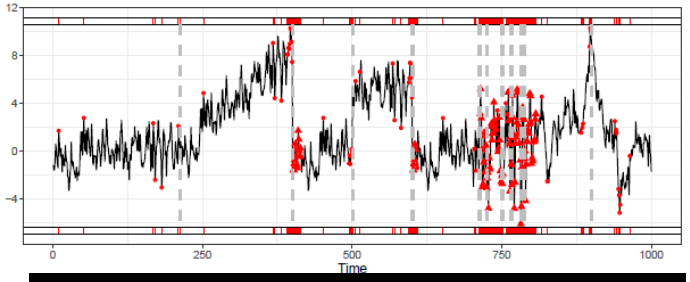
Método A: anomalias



Método B: anomalias & pontos de mudança



Método C: anomalias de volatilidade



Métodos A, B & C:
anomalias, anomalias de volatilidade e pontos de mudança

[1] R. Salles, L. Escobar, L. Baroni, R. Zorrilla, A. Ziviani, V. Kreischer, F. Delicato, P. Pires, L. Maia, et al., 2020, Um framework para integração e análise de métodos de detecção de eventos em séries temporais, In: *Anais do Simpósio Brasileiro de Banco de Dados (SBBD)*

Dimensionalidade

- Univariada e multivariada

value
13.8
13.9
14.1
13.8
13.9
13.9
14.1
14.0
14.1
14.2

(a)

time	value
1971	13.8
1972	13.9
1973	14.1
1974	13.8
1975	13.9
1976	13.9
1977	14.1
1978	14.0
1979	14.1
1980	14.2

(b)

time	global temperature	crude oil production
1971	13.8	2491
1972	13.9	2634
1973	14.1	2870
1974	13.8	2875
1975	13.9	2740
1976	13.9	2966
1977	14.1	3069
1978	14.0	3108
1979	14.1	3229
1980	14.2	3111

(c)

Granularidade

- Agregação temporal

year\month	1	2	3	4	5	6	7	8	9	10	11	12
1971	13.9	13.8	13.8	13.8	13.9	13.8	13.9	13.9	13.9	13.8	13.9	13.9
1972	13.7	13.7	14.0	14.0	13.9	14.0	14.0	14.0	13.9	14.0	14.0	14.0
1973	14.3	14.3	14.3	14.2	14.1	14.2	14.1	14.0	14.0	14.0	13.9	14.0
1974	13.8	13.7	13.9	13.9	13.9	13.8	13.9	14.0	13.9	13.8	13.8	13.8
1975	14.0	14.0	14.0	14.0	14.0	14.0	13.9	13.9	13.9	13.8	13.7	13.8
1976	13.9	13.8	13.8	13.9	13.8	13.9	13.9	13.9	13.9	13.7	13.9	14.0
1977	14.1	14.1	14.2	14.2	14.2	14.2	14.1	14.1	14.1	14.0	14.1	14.0
1978	14.1	14.1	14.1	14.0	14.0	14.0	14.0	13.9	14.0	14.0	14.1	14.0
1979	14.0	13.8	14.1	14.0	14.1	14.1	14.1	14.2	14.2	14.2	14.2	14.4
1980	14.3	14.3	14.2	14.2	14.3	14.2	14.2	14.1	14.1	14.1	14.2	14.1

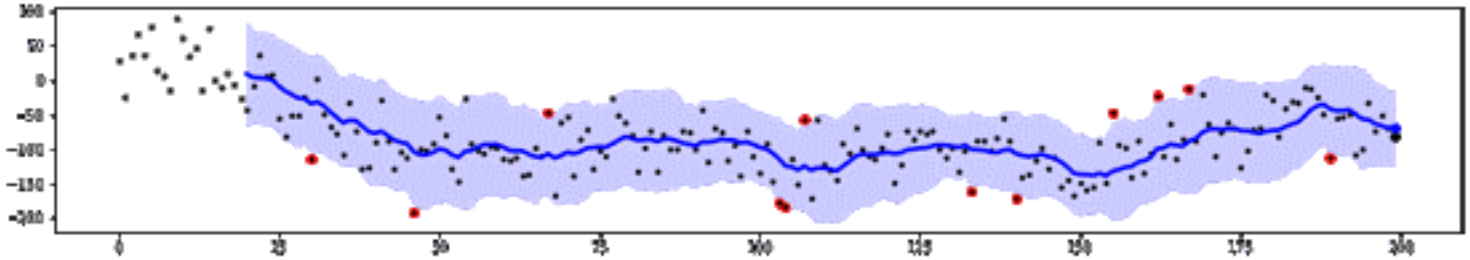
(a)

year	value
1971	13.8
1972	13.9
1973	14.1
1974	13.8
1975	13.9
1976	13.9
1977	14.1
1978	14.0
1979	14.1
1980	14.2

(b)

Detecção offline, online e predição de eventos

offline



online

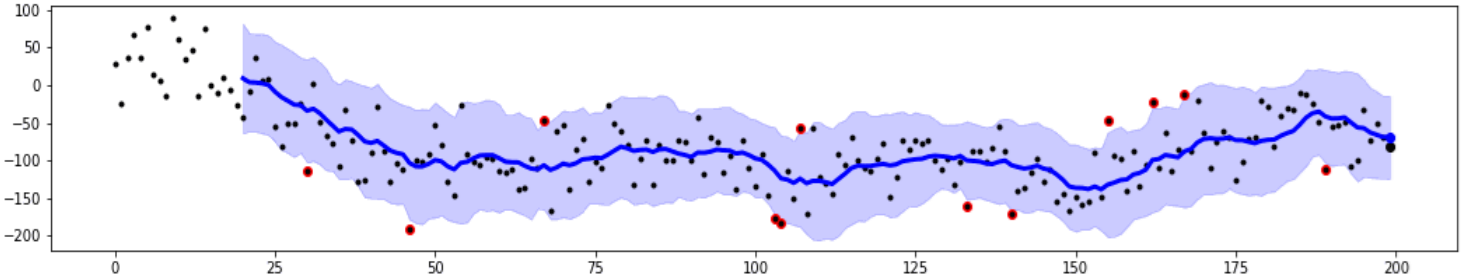


Image source: https://towardsdatascience.com/real-time-time-series-anomaly-detection-981cf1e1ca13

Estratégias de detecção (classificação)

t	x_{t-4}	x_{t-3}	x_{t-2}	x_{t-1}	x_t	$\hat{e}_t$	e_t
5	v_1	v_2	v_3	v_4	v_5	$\hat{b}_5$	b_5
6	v_2	v_3	v_4	v_5	v_6	$\hat{b}_6$	b_6
7	v_3	v_4	v_5	v_6	v_7	$\hat{b}_7$	b_7
8	v_4	v_5	v_6	v_7	v_8	$\hat{b}_8$	b_8
9	v_5	v_6	v_7	v_8	v_9	$\hat{b}_9$	b_9
10	v_6	v_7	v_8	v_9	v_{10}	$\hat{b}_{10}$	b_{10}
11	v_7	v_8	v_9	v_{10}	v_{11}	$\hat{b}_{11}$	b_{11}
12	v_8	v_9	v_{10}	v_{11}	v_{12}	$\hat{b}_{12}$	b_{12}

(a)

t	x_{t-4}	x_{t-3}	x_{t-2}	x_{t-1}	x_t	$\hat{e}_t$
13	v_9	v_{10}	v_{11}	v_{12}	v_{13}	$\hat{b}_{13}$
14	v_{10}	v_{11}	v_{12}	v_{13}	v_{14}	$\hat{b}_{14}$

(b)

Estratégias de detecção (regressão)

t	x_{t-4}	x_{t-3}	x_{t-2}	x_{t-1}	$\hat{x}_t$	x_t
5	v_1	v_2	v_3	v_4	$\hat{v}_5$	v_5
6	v_2	v_3	v_4	v_5	$\hat{v}_6$	v_6
7	v_3	v_4	v_5	v_6	$\hat{v}_7$	v_7
8	v_4	v_5	v_6	v_7	$\hat{v}_8$	v_8
9	v_5	v_6	v_7	v_8	$\hat{v}_9$	v_9
10	v_6	v_7	v_8	v_9	$\hat{v}_{10}$	v_{10}
11	v_7	v_8	v_9	v_{10}	$\hat{v}_{11}$	v_{11}
12	v_8	v_9	v_{10}	v_{11}	$\hat{v}_{12}$	v_{12}
13	v_9	v_{10}	v_{11}	v_{12}	$\hat{v}_{13}$	v_{13}
14	v_{10}	v_{11}	v_{12}	v_{13}	$\hat{v}_{14}$	v_{14}

(a)

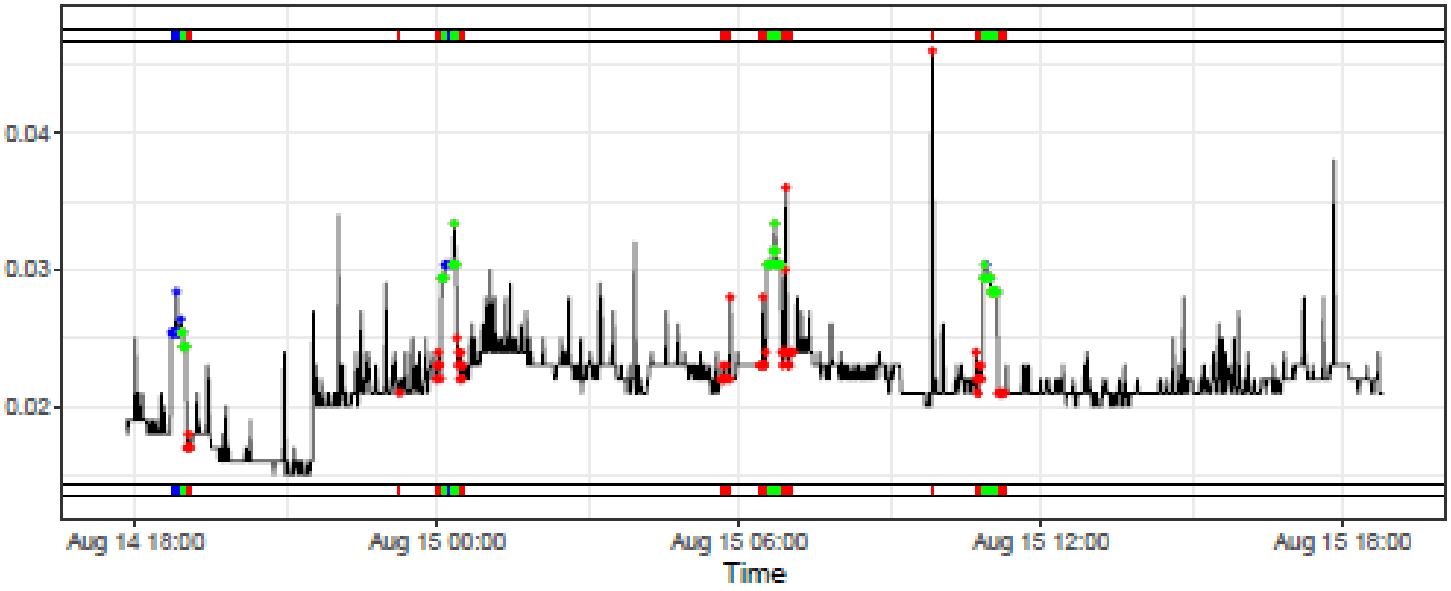
Estratégias de detecção (agrupamento)

t	x_{t-4}	x_{t-3}	x_{t-2}	x_{t-1}	x_t	$\ddot{r}_c$	d_t
5	v_1	v_2	v_3	v_4	v_5	$\ddot{r}_1$	d_5
6	v_2	v_3	v_4	v_5	v_6	$\ddot{r}_1$	d_6
7	v_3	v_4	v_5	v_6	v_7	$\ddot{r}_1$	d_7
8	v_4	v_5	v_6	v_7	v_8	$\ddot{r}_2$	d_8
9	v_5	v_6	v_7	v_8	v_9	$\ddot{r}_2$	d_9
10	v_6	v_7	v_8	v_9	v_{10}	$\ddot{r}_1$	d_{10}
11	v_7	v_8	v_9	v_{10}	v_{11}	$\ddot{r}_1$	d_{11}
12	v_8	v_9	v_{10}	v_{11}	v_{12}	$\ddot{r}_2$	d_{12}
13	v_9	v_{10}	v_{11}	v_{12}	v_{13}	$\ddot{r}_2$	d_{13}
14	v_{10}	v_{11}	v_{12}	v_{13}	v_{14}	$\ddot{r}_2$	d_{14}

$\ddot{r}_t$	x_{t-4}	x_{t-3}	x_{t-2}	x_{t-1}	x_t
$\ddot{r}_1$	$\ddot{v}_{1,4}$	$\ddot{v}_{1,3}$	$\ddot{v}_{1,2}$	$\ddot{v}_{1,1}$	$\ddot{v}_{1,0}$
$\ddot{r}_2$	$\ddot{v}_{2,4}$	$\ddot{v}_{2,3}$	$\ddot{v}_{2,2}$	$\ddot{v}_{2,1}$	$\ddot{v}_{2,0}$

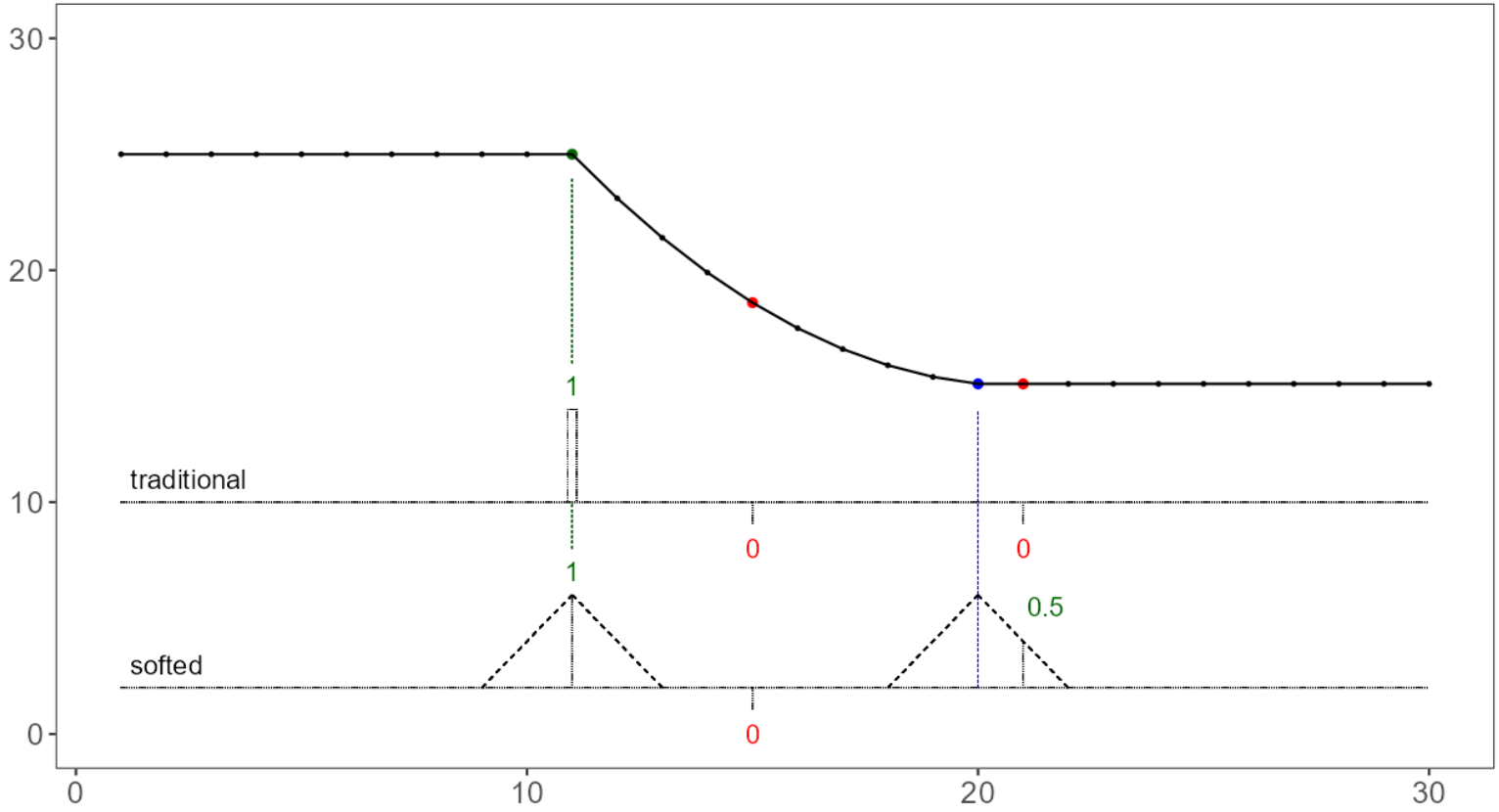
Métricas para avaliar detecção de eventos

- Métricas de classificação clássicas:
  - acurácia $= \frac{TP+TN}{All}$
  - precisão $= \frac{TP}{TP+FP}$
  - revocação $= \frac{TP}{TP+FN}$
  - $F_1 = \frac{2 \cdot \text{precisão} \cdot \text{revocação}}{\text{precisão} + \text{revocação}}$



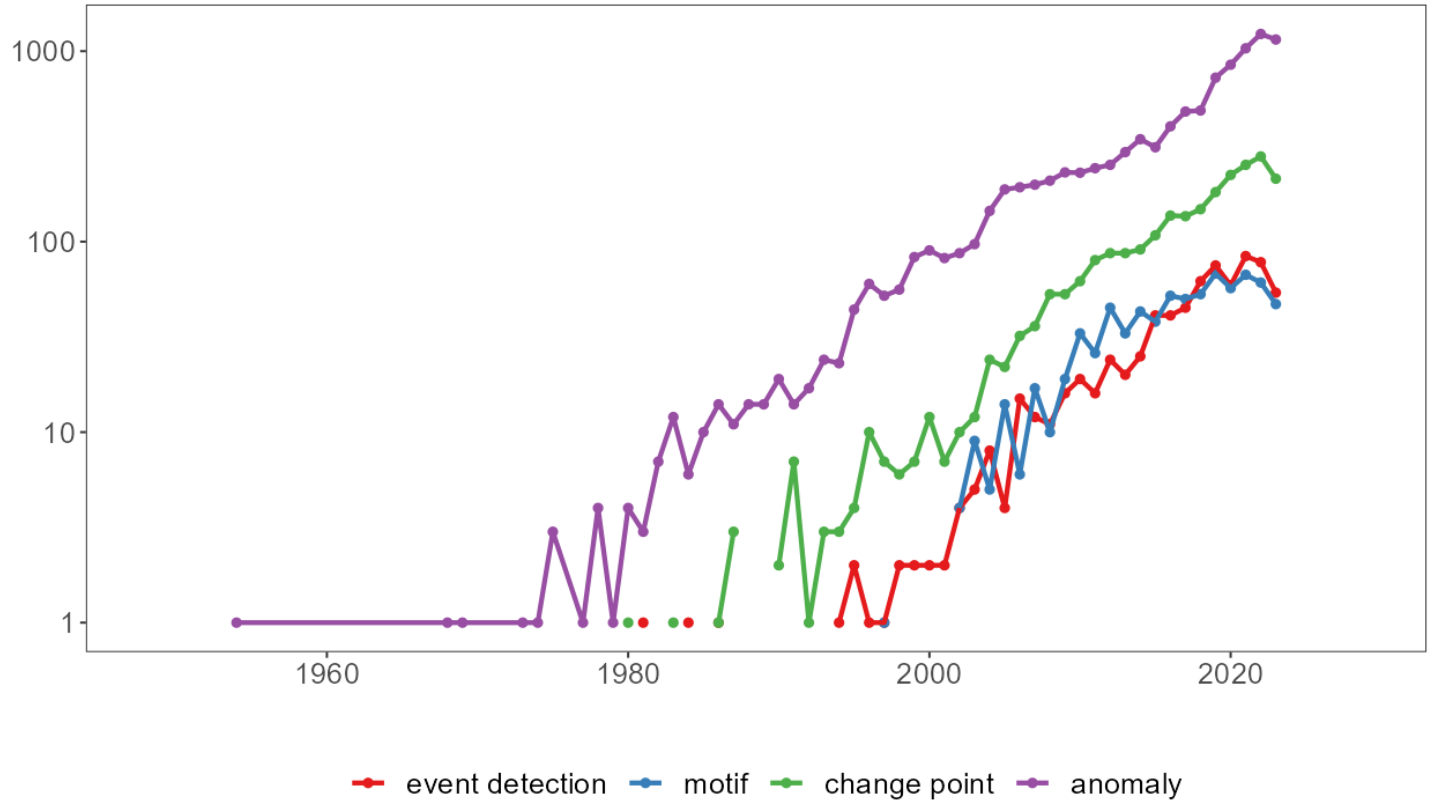
Desafios nas métricas de avaliação

- Métodos de pontuação tradicionais, como precisão e revocação, não são suficientes para avaliar o desempenho da detecção de eventos online
- Eles não incorporam o tempo e não recompensam a detecções próximas
  - Verdadeiros positivos são recompensados
- Todos os outros resultados são "severamente" e igualmente punidos



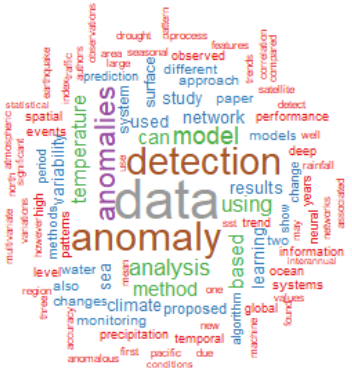
Visão geral da pesquisa

Pesquisa


1000
100
10
1
1960
1980
2000
2020
event detection
motif
change point
anomaly

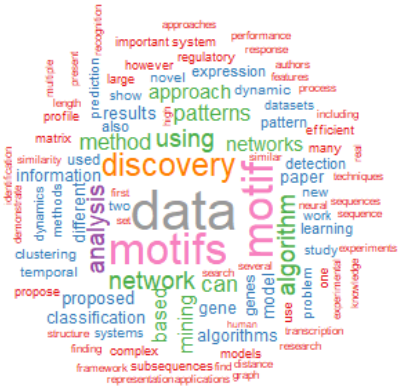
Palavras-chaves

(a)



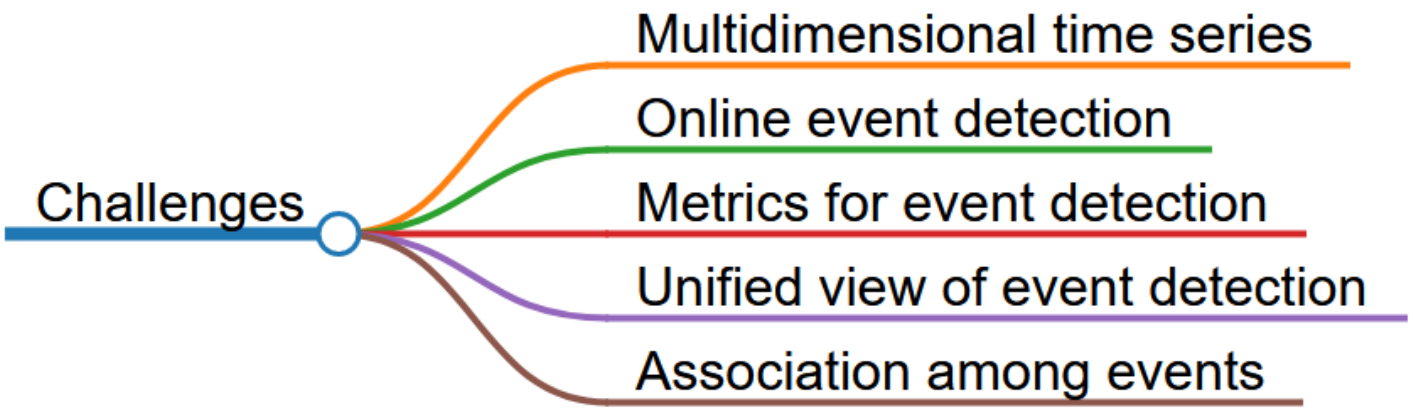
(b)

(c)



(d)

Desafios


Challenges
Multidimensional time series
Online event detection
Metrics for event detection
Unified view of event detection
Association among events

Materiais de apoio

- Curso de Análise de Dados
  - https://eic.cefet-rj.br/~eogasawara/analise-de-dados/
- Curso de Mineração de Dados
  - Slides e vídeos em: https://eic.cefet-rj.br/~eogasawara/data-mining/
- Tutorial de R
  - Vídeos em: https://eic.cefet-rj.br/~eogasawara/tutorial-r/