I Jornada em Computação

UFRRJ
UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO

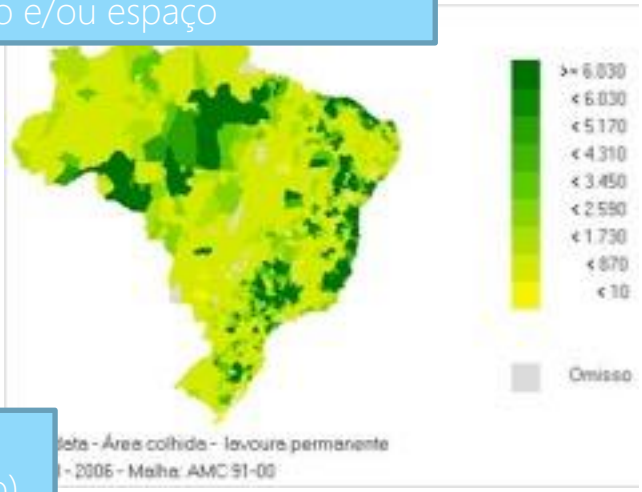
Desafios e oportunidades de pesquisa em Ciência de Dados em contextos não-estacionários

CEFET/RJ

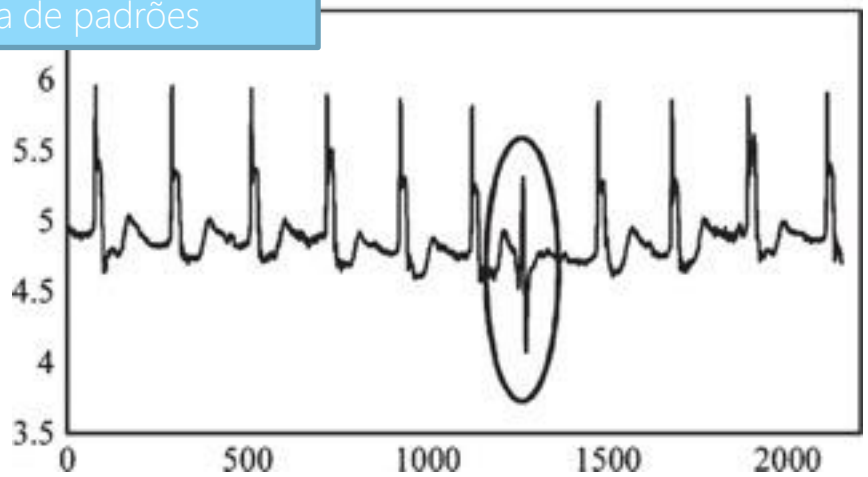
Eduardo Ogasawara
eogasawara@ieee.org
http://eic.cefet-rj.br/~eogasawara

Contexto

Muitos fenômenos são modelados por series temporais, series espaço-temporais ou variam ao longo do tempo e/ou espaço



Tomada de decisão
Predição (classificação / regressão)
Descoberta de padrões

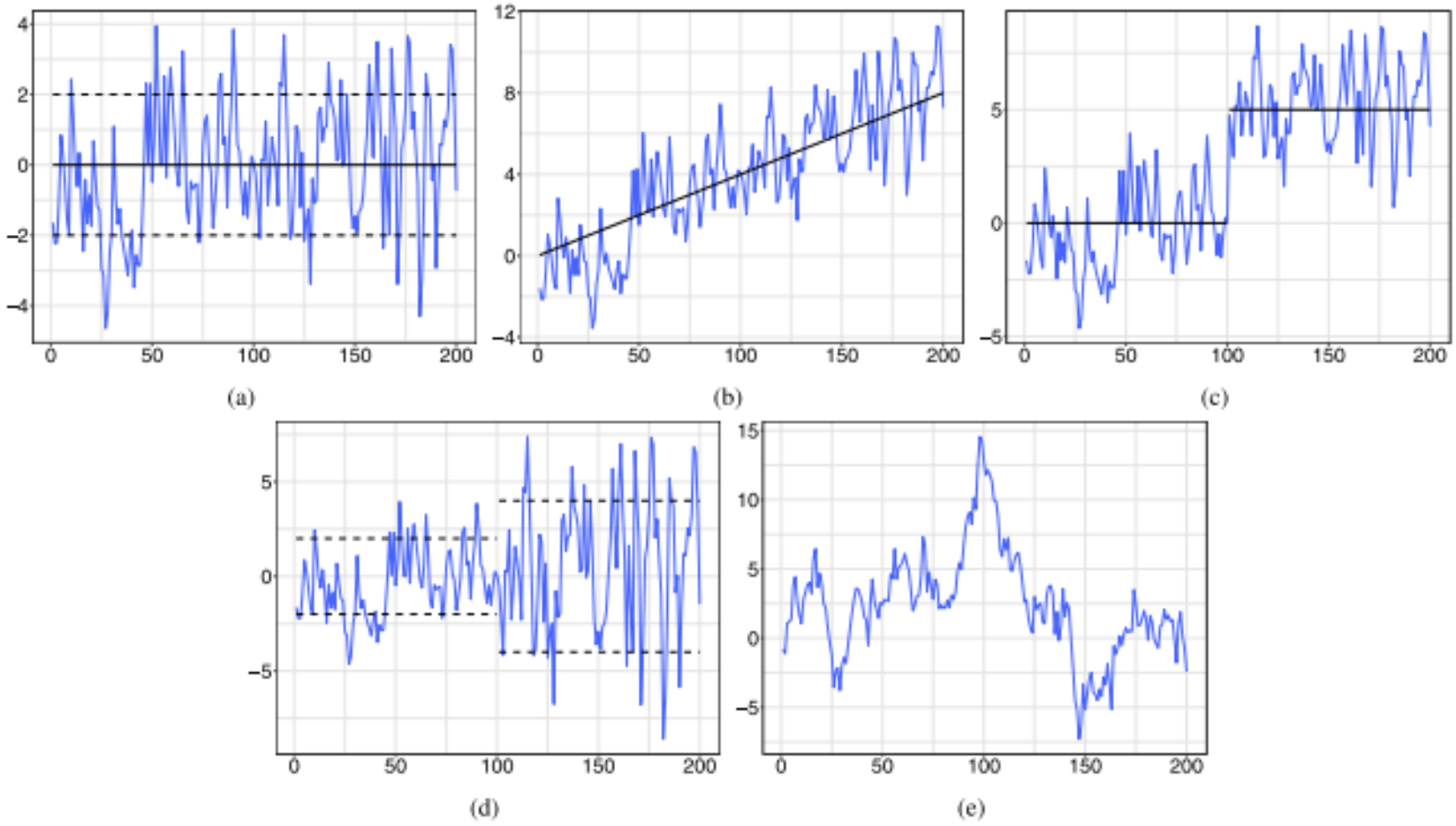


Não-estacionariedade

- Estacionariedade
  - Dataset D
  - Amostras D_s
  - Propriedades estatísticas em D_s não variam com o tempo
    - Séries temporais: média, variância e covariância
- Não-estacionariedade
- Métodos de data analytics
  - A grande maioria dos métodos assumem "implicitamente" estacionariedade

[1] R.H. Shumway e D.S. Stoffer, 2017, Time Series Analysis and Its Applications: With R Examples. 4 ed. New York, NY, Springer.

Tipos de não-estacionariedade



R. Salles, K. Belloze, F. Porto, P. H. Gonzalez, e E. Ogasawara, "Nonstationary time series transformation methods: An experimental review", *Knowledge-Based Systems*, nov. 2018.

Ciência de Dados

- Análise de Dados (*Data Analytics)*
  - Descritiva
  - Preditiva
  - Prescritiva
- A grande maioria dos métodos assumem "implicitamente" estacionariedade

As múltiplas faces da não-estacionariedade

- Crítica de Lucas
  - visão econômica
- Detecção de eventos
- Dilema da plasticidade e estabilidade
  - visão de aprendizado de máquina
- Mudança de conceito (*concept drift)*
  - visão de mineração de dados
- Padrões emergentes (*emerging patterns)*
  - visão de banco de dados

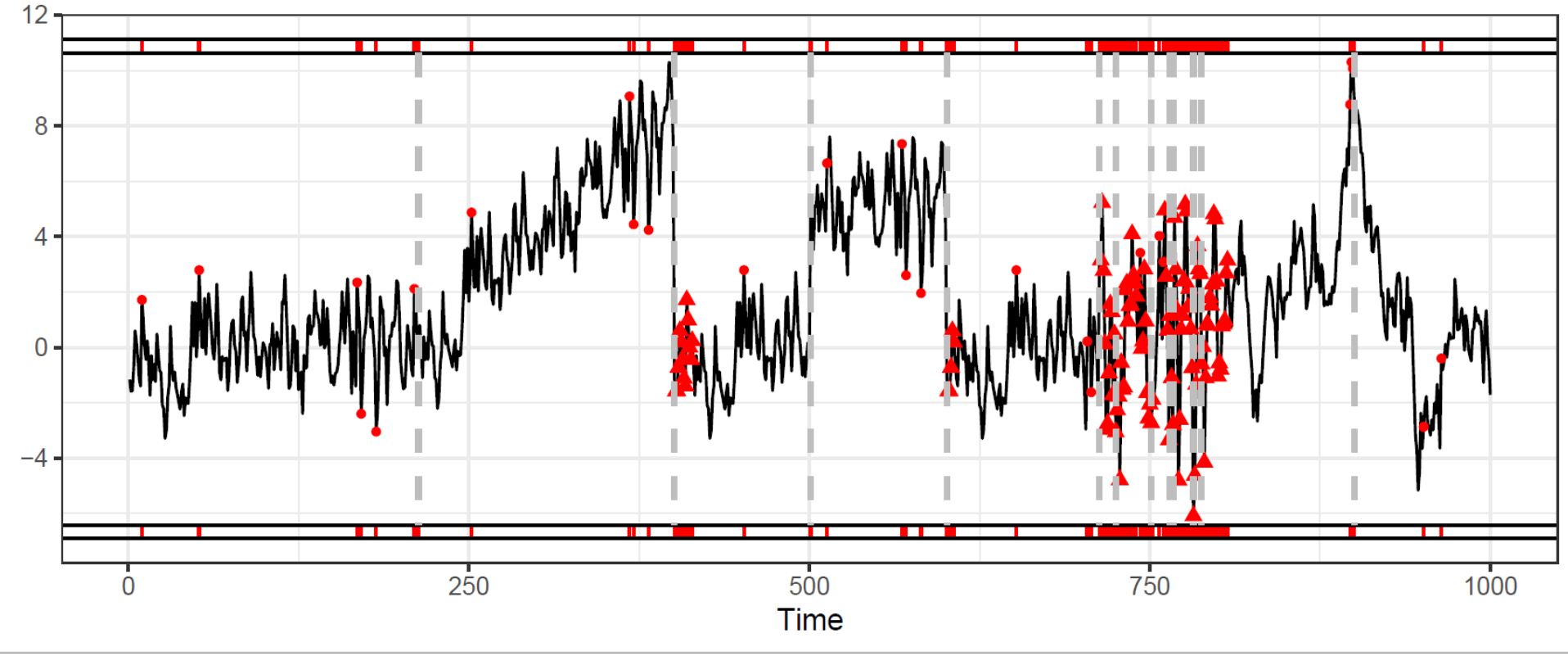
Crítica de Lucas

- "Dado que a estrutura de um modelo econométrico consiste em regras de decisão ótimas dos agentes econômicos, e que as regras de decisão ótimas variam sistematicamente com as mudanças na estrutura das séries relevantes para o decisor, conclui-se que qualquer mudança na política sistematicamente irá alterar a estrutura dos modelos econométricos"

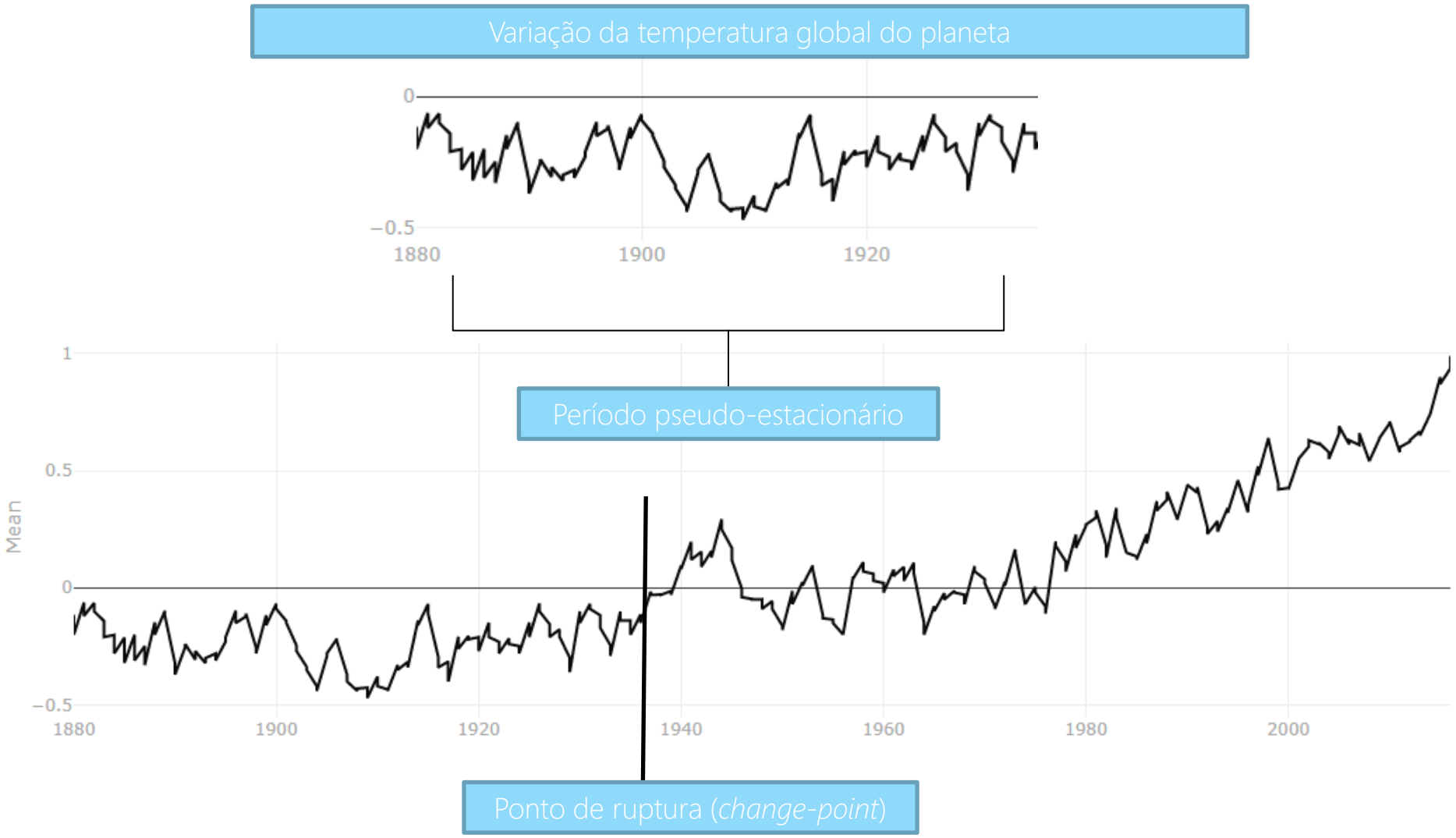
[1] D. Gujarati, 2002, *Basic Econometrics*. 4 ed. Boston; Montreal, McGraw-Hill/Irwin.

Event detection

- Events
  - Anomalies (trends and volatility)
  - Change points



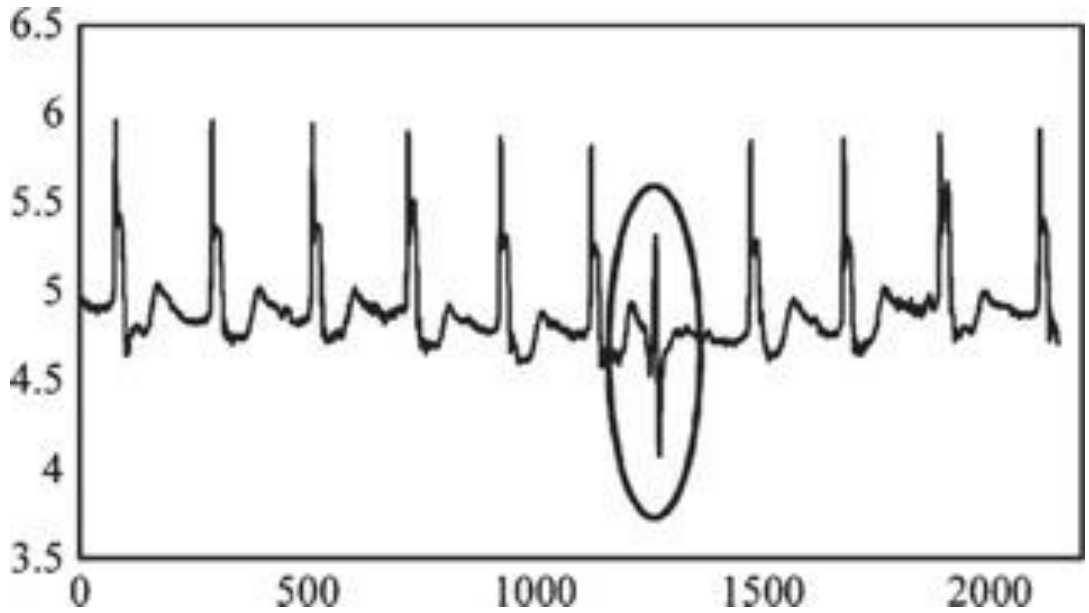
Pseudo-estacionariedade / Change points



[1] V. Guralnik e J. Srivastava, 1999, Event Detection from Time Series Data, In: *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 33–42.
[2] World Global Temperature, https://datahub.io/core/global-temp
[3] J.-I. Takeuchi e K. Yamanishi, 2006, A unifying framework for detecting outliers and change points from time series, IEEE Transactions on Knowledge and Data Engineering, v. 18, n. 4, p. 482–492.

Anomaly

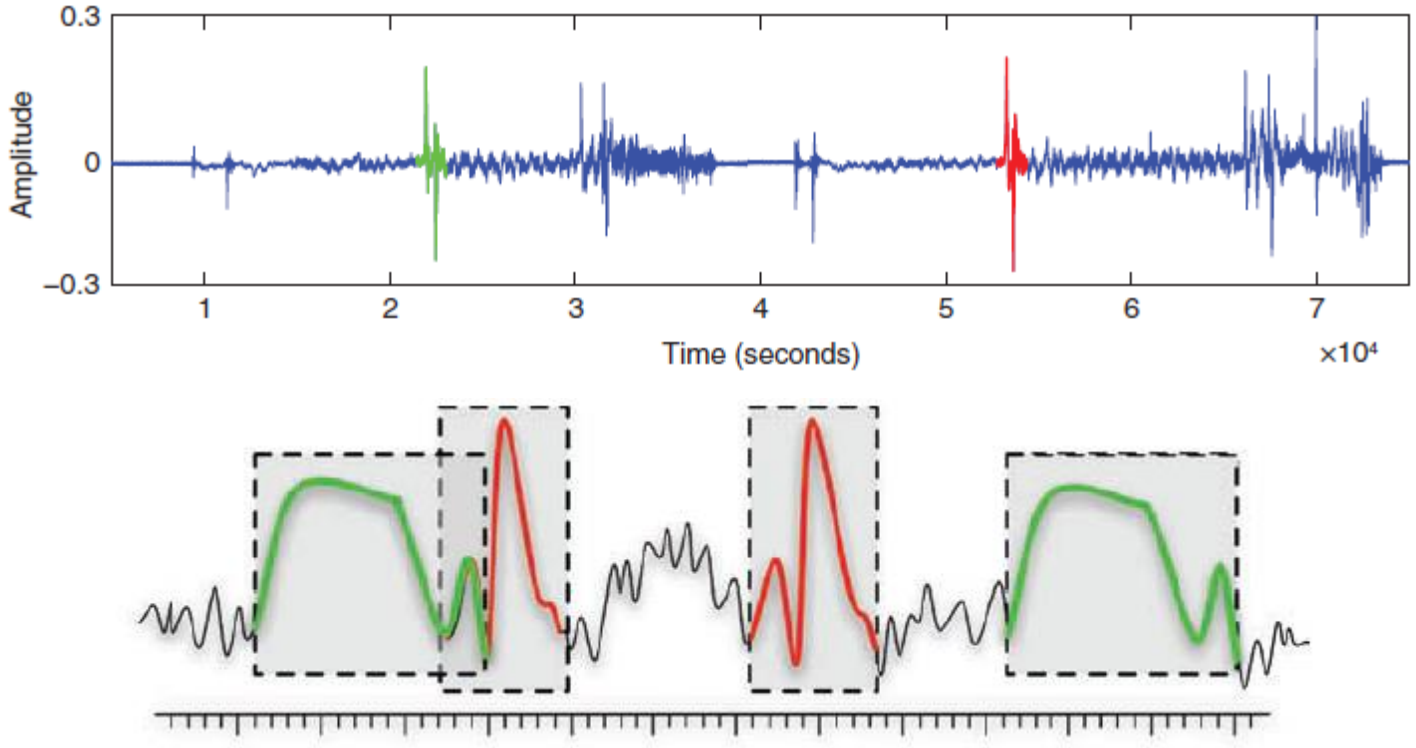
- Anomaly is a pattern that do not conform to expected behavior [1]
- Anomaly detection refers to the problem of finding these patterns



[1] V. Chandola, A. Banerjee, e V. Kumar, 2009, Anomaly detection: A survey, *ACM Computing Surveys*, v. 41, n. 3

Motifs

- A pattern (unknown) that occurs a significant number of times in time series [1,2,3]

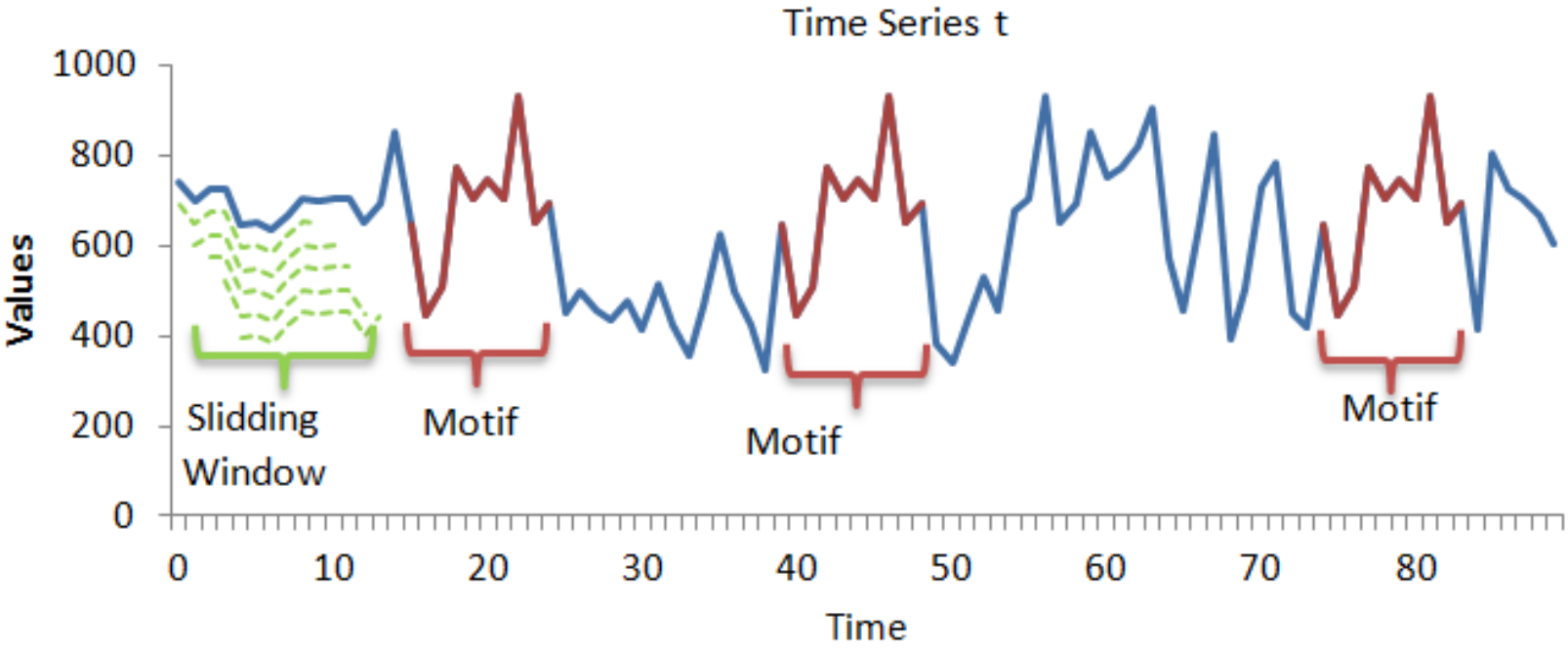


[1] P. Patel, E. Keogh, J. Lin, and S. Lonardi, "Mining motifs in massive time series databases," in Proceedings - IEEE International Conference on Data Mining, ICDM, 2002, pp. 370–377
[2] A. Mueen, "Time series motif discovery: Dimensions and applications," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 4, no. 2, pp. 152–159, 2014
[3] S. Torkamani and V. Lohweg, "Survey on time series motif discovery," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 7, no. 2, 2017.

Motif in time series

A sequence $s = <w_1, w_2, \cdots, w_k>$ is **included** in time series $t = <v_1, v_2, \cdots, v_n>$ if there exist integers $i_1 < i_2 < \cdots < i_k$ such that $w_1 = v_{i_1}, w_2 = v_{i_2}, \cdots, w_k = v_{i_k}$.

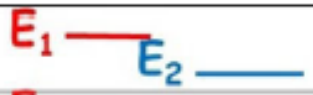
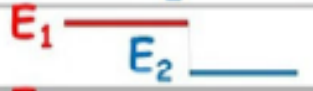
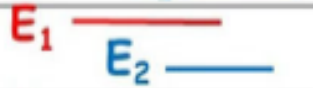
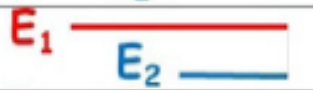
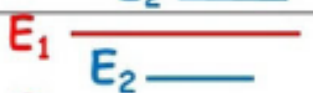
Given a time series t and sequence q, q is a motif for t with support σ iff q is included in t at least σ times. Formally, given time series t and q, such that $W = sw(t, |q|) \iff \exists R \subseteq W | \forall w_i \in R, w_i = q \wedge |R| \geq \sigma$.



What is a motif in spatial-time series?
How to find motifs in spatial-time series?
How to do it in non-stationarity?

Association of events

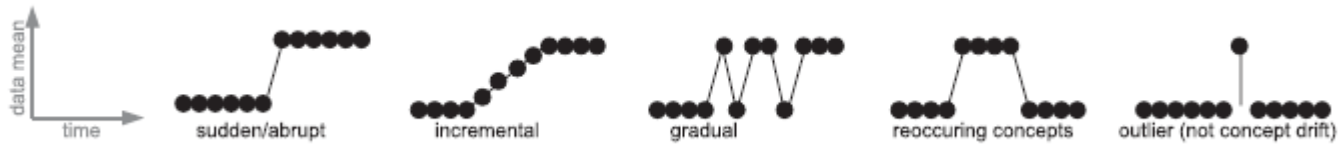
- If you can label or characterize events
  -

	E_1 before E_2	E_2 after E_1
	E_1 meets E_2	E_2 is-met-by E_1
	E_1 overlaps E_2	E_2 is-overlapped-by E_1
	E_1 is-finished-by E_2	E_2 finishes E_1
	E_1 contains E_2	E_2 during E_1
	E_1 starts E_2	E_2 is-started-by E_1
	E_1 equals E_2	E_2 equals E_1

[1] I. Batal, G.F. Cooper, D. Fradkin, Jr. Harrison J., F. Moerchen, e M. Hauskrecht, 2016, An efficient pattern mining approach for event detection in multivariate temporal data, *Knowledge and Information Systems*, v. 46, n. 1, p. 115–150.

Dilema da Plasticidade e Estabilidade

- Redes neurais são conhecidas pela adaptabilidade
  - Capacidade de atualizar os pesos em função de alterações no ambiente
  - Treinamento incremental
    - Alteração dos pesos sinápticos
- Sistemas adaptativos visam abordar não-estacionariedade
  - Buscando-se robustez, adota-se adaptabilidade
  - Maior adaptabilidade, mais suscetível a situações espúrias, menor robustez
  - Dilema: encontrar o tempo certo para se adaptar



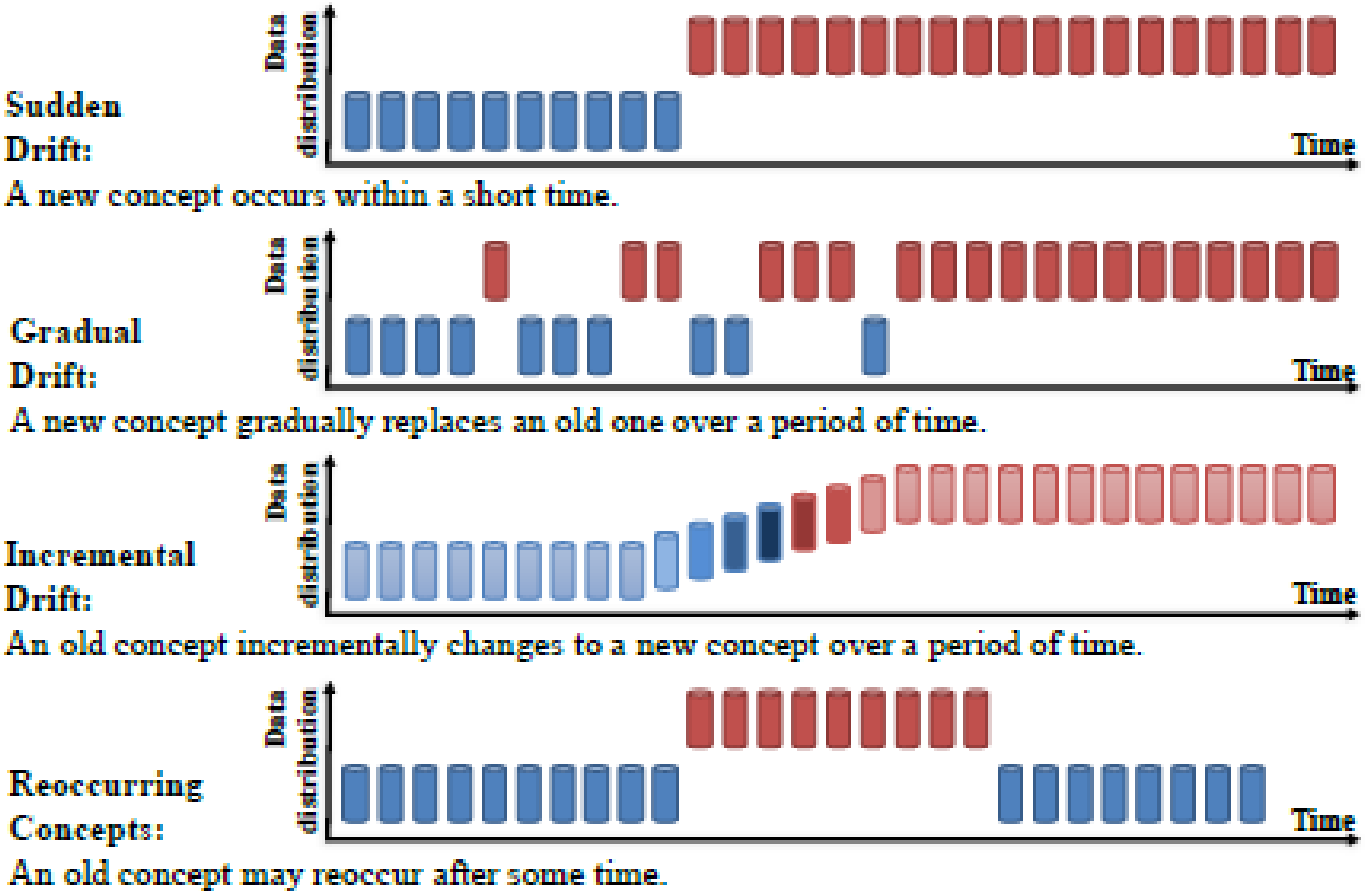
[1] S.O. Haykin, 2008, *Neural Networks and Learning Machines*. 3 ed. New York, Prentice Hall.
[2] Grossberg, S., 1988. Neural Networks and Natural Intelligence, Cambridge, MA: MIT Press.
[3] G. Ditzler, M. Roveri, C. Alippi, e R. Polikar, 2015, Learning in Nonstationary Environments: A Survey, *IEEE Computational Intelligence Magazine*, v. 10, n. 4, p. 12–25.

Concept Drift

- Aprendizado no contexto de distribuições não-estacionárias
  - Aprendizado é feito em lotes (*batches*)
  - *Data streams* (objetos com *timestamps*)
  - Definições
    - $\mathrm{P}(Y)$ probabilidade da variável dependente (rótulo)
    - $\mathrm{P}(X)$ probabilidade das variáveis independentes (objetos)
    - $\mathrm{P}(X, Y)$ probabilidade conjunta dos objetos e rótulo
    - $\mathrm{P}(Y|X)$ distribuição provável do rótulo para objeto
  - Concept = $\mathrm{P}(X, Y) = \mathrm{P}(\chi)$
  - Drift = $P_t(\chi) \neq P_u(\chi)$

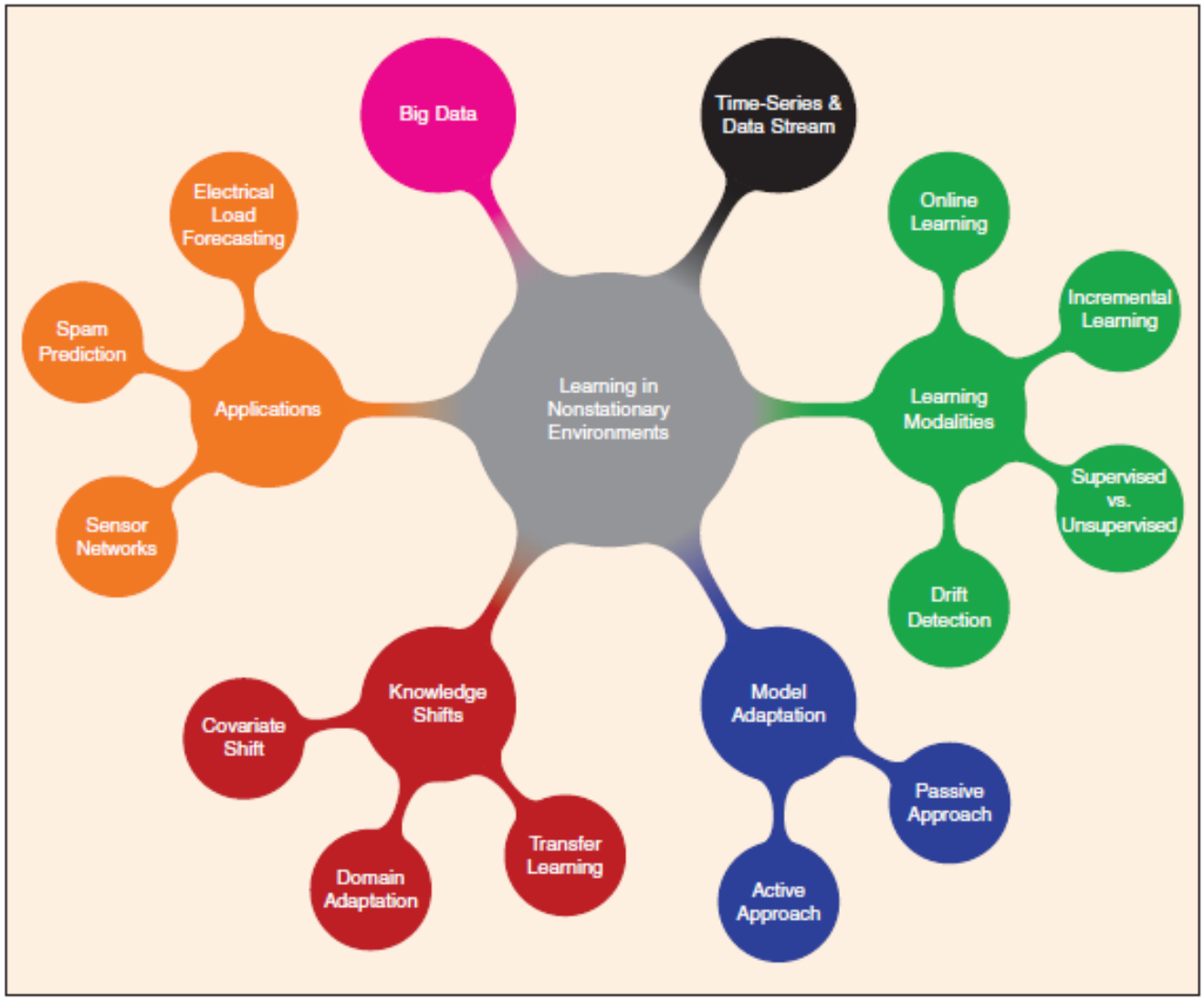
[1] G.I. Webb, R. Hyde, H. Cao, H.L. Nguyen, e F. Petitjean, 2016, Characterizing concept drift, *Data Mining and Knowledge Discovery*, v. 30, n. 4, p. 964–994.

Tipos de Concept Drift



[3] 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.

Taxonomia de não-estacionariedade



[3] 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.

Magnitude e Real/Virtual Concept Drift

- Magnitude do *Concept Drift*: D(t, u)
- *Real Concept Drift*
  - $P_t(Y|X) \neq P_u(Y|X)$ e $P_t(X) = P_u(X)$
- *Virtual Concept Drift*
  - $P_t(Y|X) = P_u(Y|X)$ e $P_t(X) \neq P_u(X)$

[1] G.I. Webb, R. Hyde, H. Cao, H.L. Nguyen, e F. Petitjean, 2016, Characterizing concept drift, *Data Mining and Knowledge Discovery*, v. 30, n. 4, p. 964–994.

Emerging patterns

- Padrões emergentes são coleções de itens cuja frequência muda de um *dataset* (*batch)* para outro
- *Datasets* D_t(anterior) e D_u (próximo)
  - Crescimento para itens χ: $\rho(\chi)$
  - $$\rho(\chi) = \begin{cases} \infty, support_t(i) = 0 \\ 0, support_t(\chi) = support_u(\chi) = 0 \\ \frac{support_u(\chi)}{support_t(\chi)}, otherwise \end{cases}$$
  - Dado um limite σ, um padrão χ é emergente se $\rho(\chi) \geq \sigma$

[1] G. Dong e J. Li, 1999, Efficient Mining of Emerging Patterns: Discovering Trends and Differences, In: *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 43–52
[2] K. Ramamohanarao e J. Bailey, 2004, Emerging Patterns: Mining and Applications, In: *Proceedings of International Conference on Intelligent Sensing and Information Processing, ICISIP 2004*, p. 409–414.

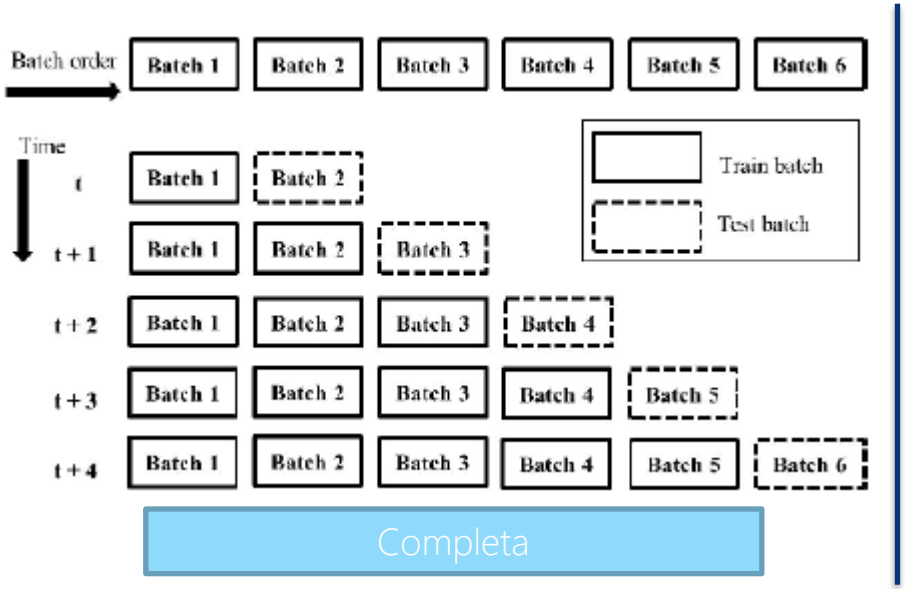
Abordagens para não-estacionariedade

- Gerência de memória
- Adaptabilidade
- Transformações

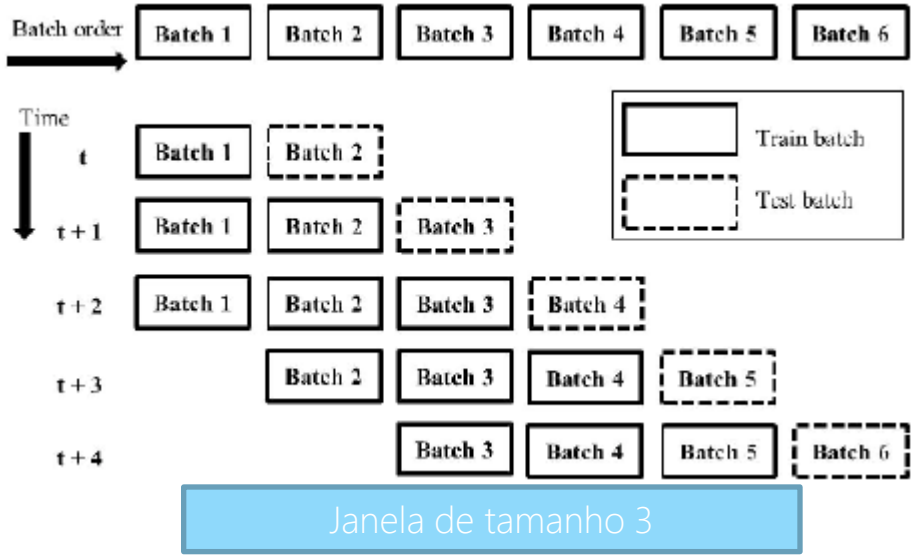
[1] J. Gama, I. Zliobaite, A. Bifet, M. Pechenizkiy, e A. Bouchachia, 2014, A survey on concept drift adaptation, *ACM Computing Surveys*, v. 46, n. 4
[2] A.M. García-Vico, C.J. Carmona, D. Martín, M. García-Borroto, e M.J. del Jesus, 2018, An overview of emerging pattern mining in supervised descriptive rule discovery: taxonomy, empirical study, trends, and prospects, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, v. 8, n. 1
[3] R. Salles, K. Belloze, F. Porto, P. H. Gonzalez, e E. Ogasawara, "Nonstationary time series transformation methods: An experimental review", *Knowledge-Based Systems*, nov. 2018.

Gerência de memória

- Processo
  - Testa-se no último batch (previsão)
  - Incorpora-se último batch no treino
- Memória
  - Completa
  - Sem memória
  - Janelas deslizantes



Completa



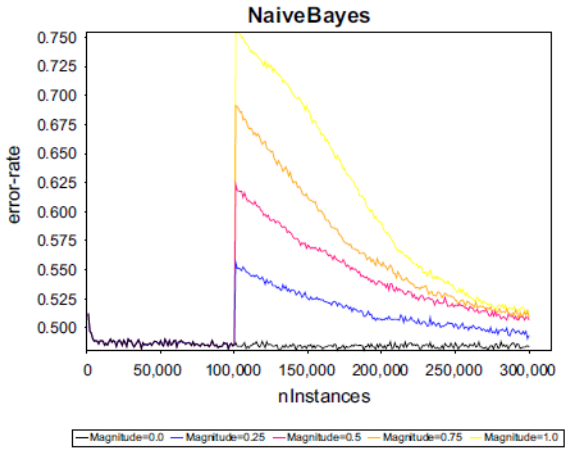
Janela de tamanho 3

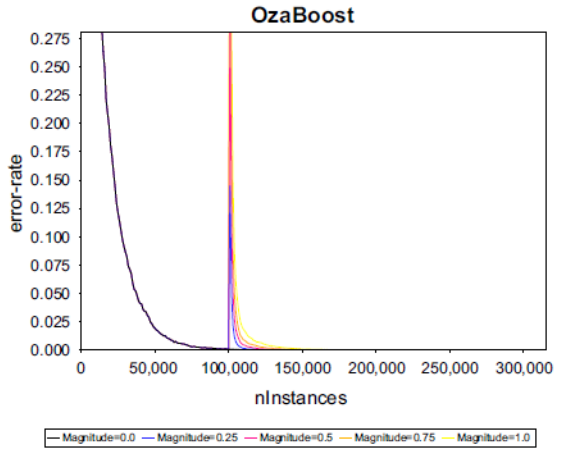
[1] J. Gama, I. Zliobaite, A. Bifet, M. Pechenizkiy, e A. Bouchachia, 2014, A survey on concept drift adaptation, *ACM Computing Surveys*, v. 46, n. 4

[2] A.M. García-Vico, C.J. Carmona, D. Martín, M. García-Borroto, e M.J. del Jesus, 2018, An overview of emerging pattern mining in supervised descriptive rule discovery: taxonomy, empirical study, trends, and prospects, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, v. 8, n. 1

Adaptabilidade

- Detecção de *drift*
  - Ativa
  - Passiva
- Aprendizado
  - Incremental
  - Não-incremental
- Modelos
  - Singular
  - Ensemble



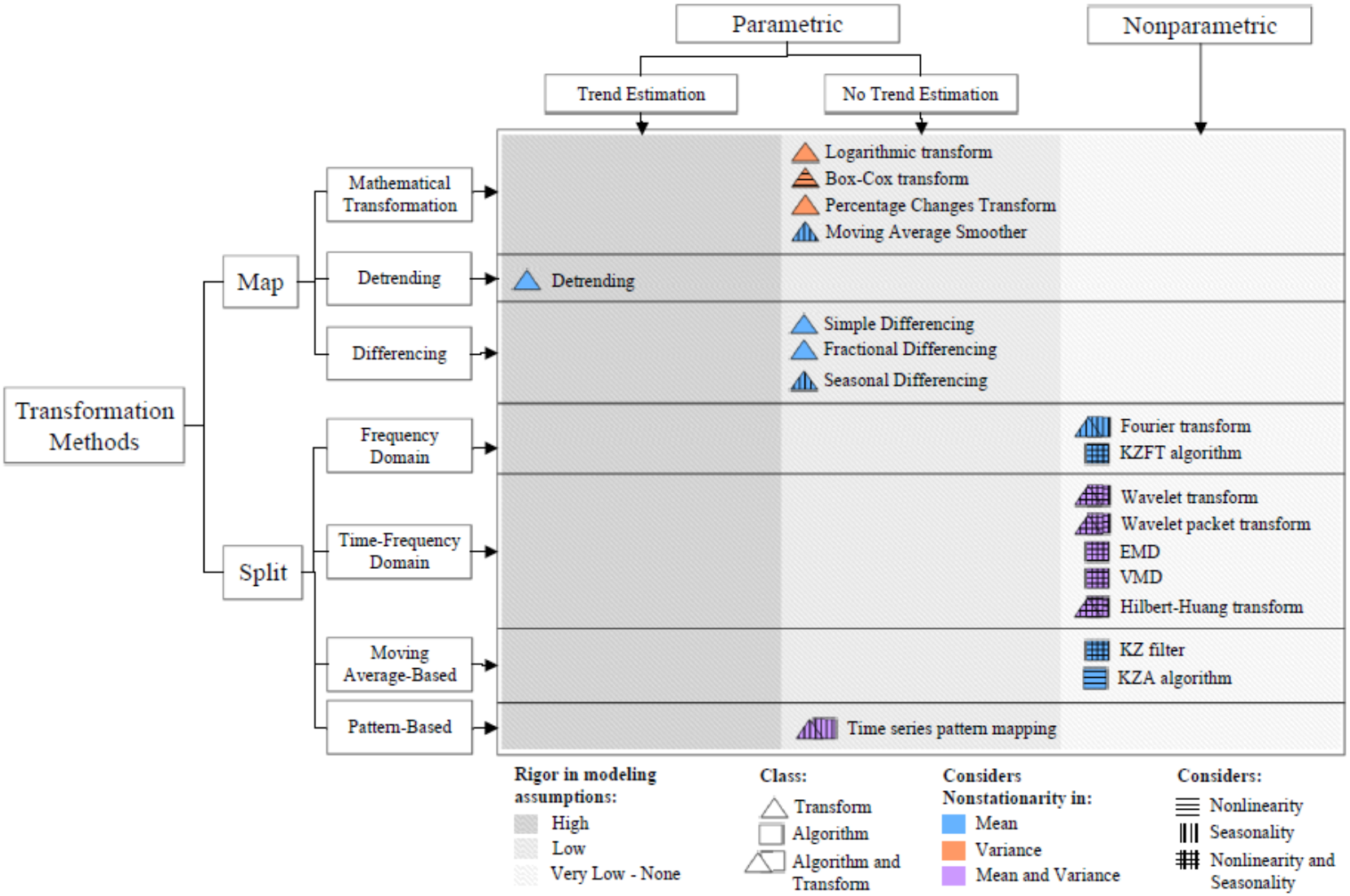


[1] J. Gama, I. Zliobaite, A. Bifet, M. Pechenizkiy, e A. Bouchachia, 2014, A survey on concept drift adaptation, *ACM Computing Surveys*, v. 46, n. 4

[2] A.M. García-Vico, C.J. Carmona, D. Martín, M. García-Borroto, e M.J. del Jesus, 2018, An overview of emerging pattern mining in supervised descriptive rule discovery: taxonomy, empirical study, trends, and prospects, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, v. 8, n. 1

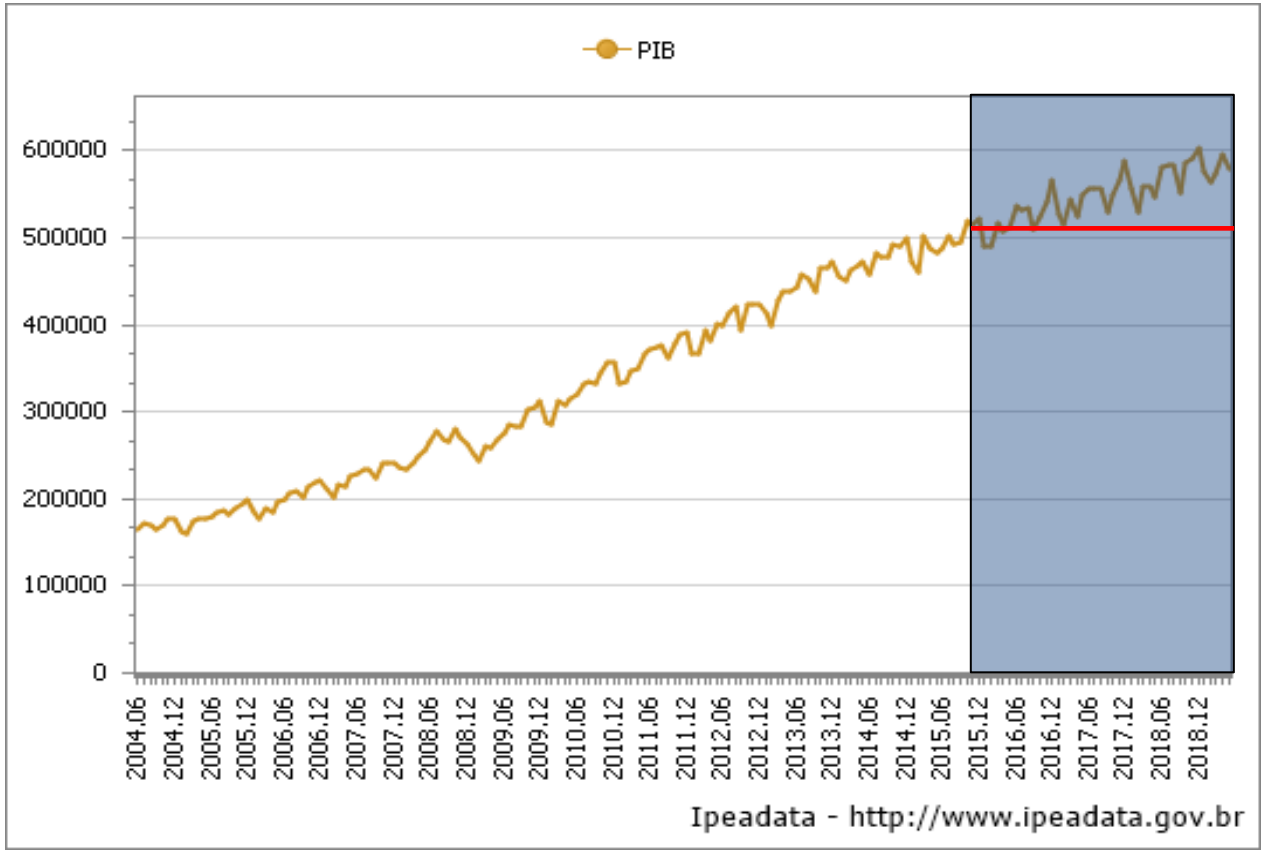
[3] G.I. Webb, R. Hyde, H. Cao, H.L. Nguyen, e F. Petitjean, 2016, Characterizing concept drift, *Data Mining and Knowledge Discovery*, v. 30, n. 4, p. 964–994.

Transformações



R. Salles, K. Belloze, F. Porto, P. H. Gonzalez, e E. Ogasawara, "Nonstationary time series transformation methods: An experimental review", *Knowledge-Based Systems*, nov. 2018.

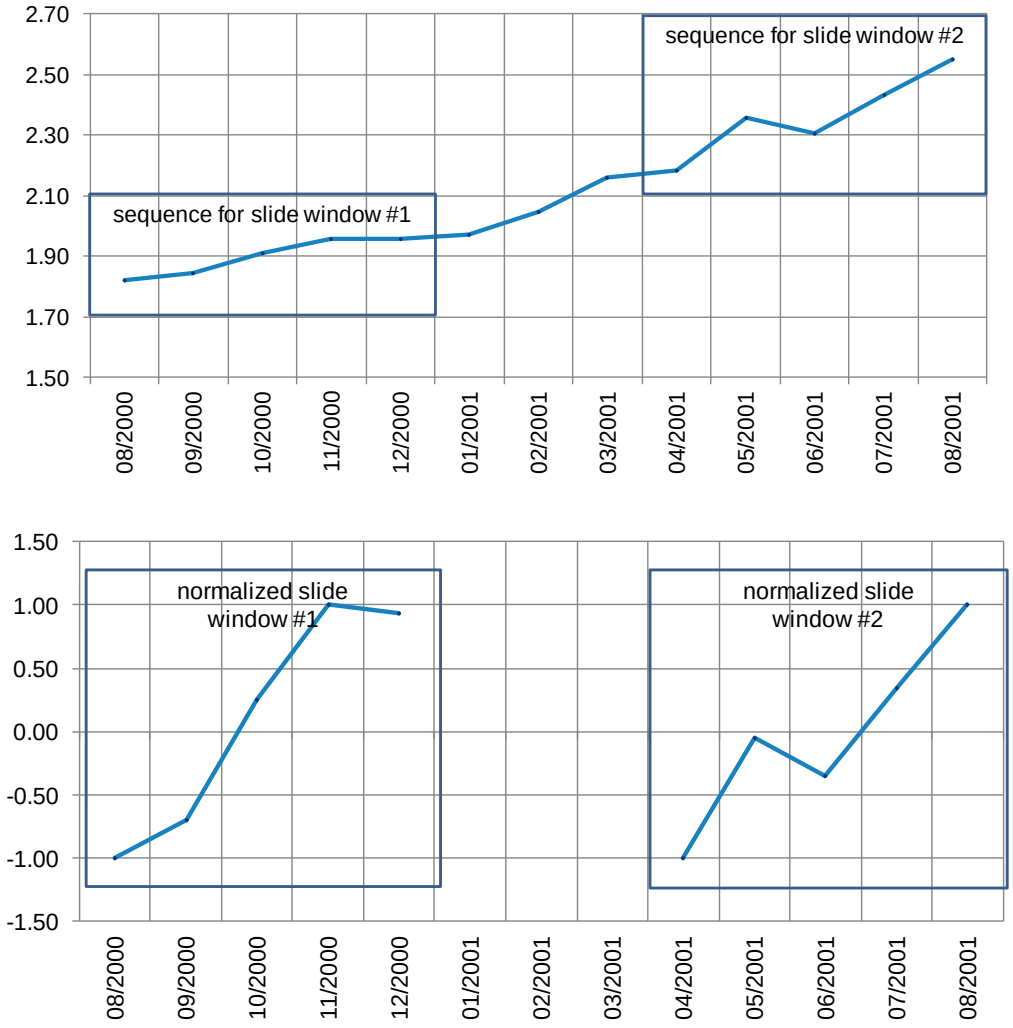
Predição de séries temporais



Normalização (min/max) global

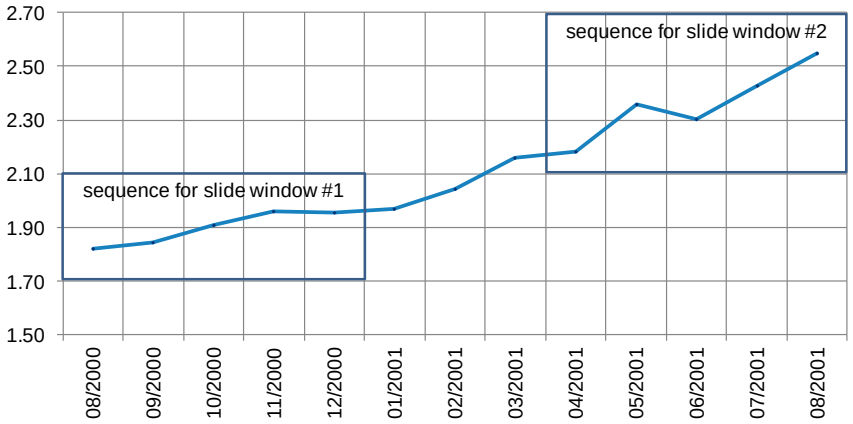
[1] E. Ogasawara, L.C. Martinez, D. De Oliveira, G. Zimbrão, G.L. Pappa, e M. Mattoso, 2010, Adaptive Normalization: A novel data normalization approach for non-stationary time series, In: *Proceedings of the International Joint Conference on Neural Networks*

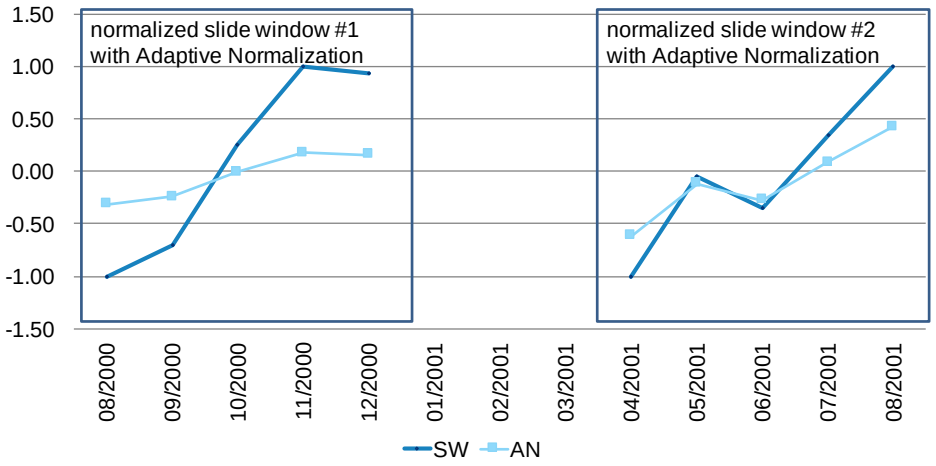
Problemas de normalização usando janelas deslizantes



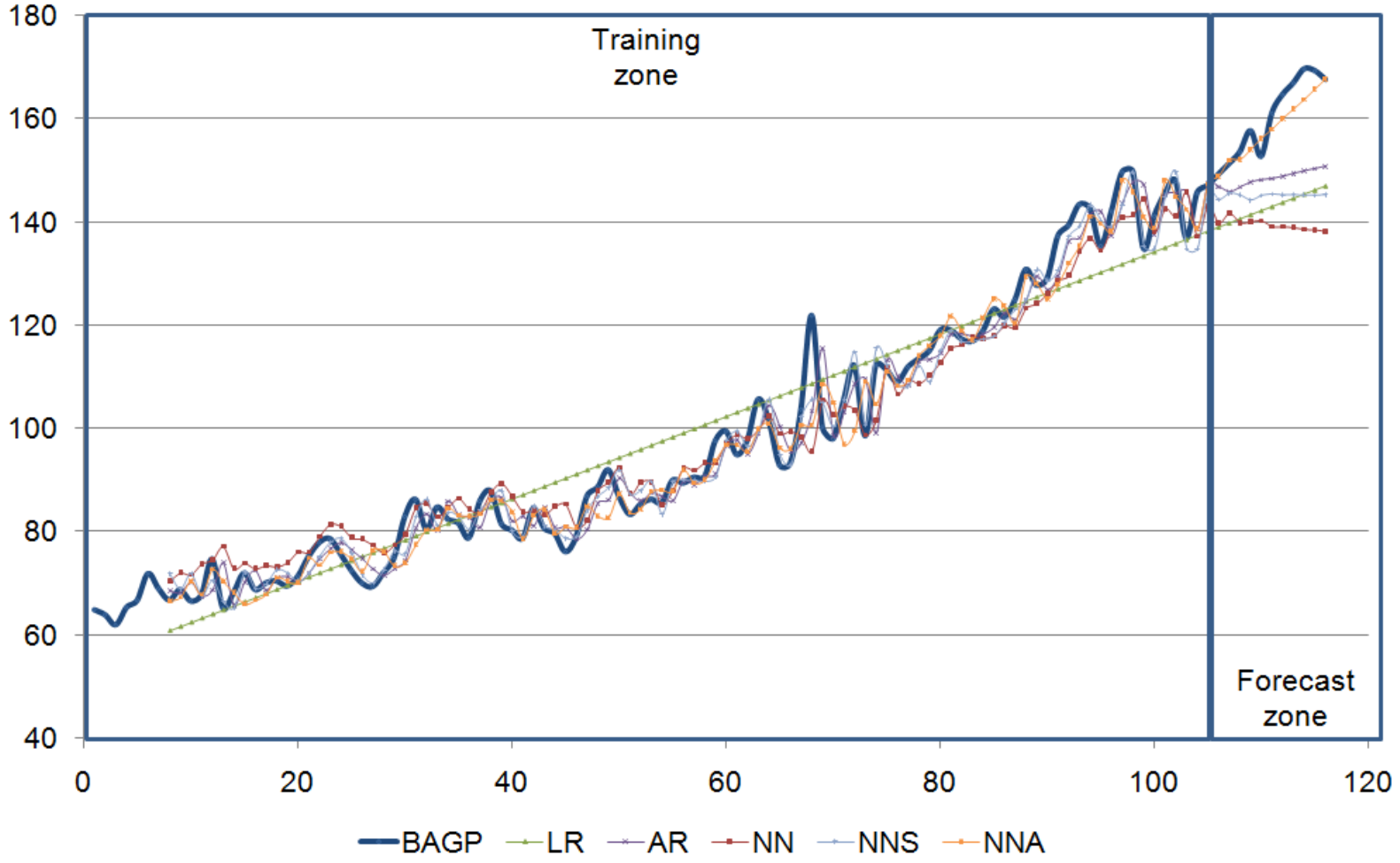
Monthly average exchange rate of U.S. Dollar to Brazilian Real
normalized by sliding window technique from aug/2000 to dec/2000 and from apr/2001 to

Intuição da normalização adaptativa

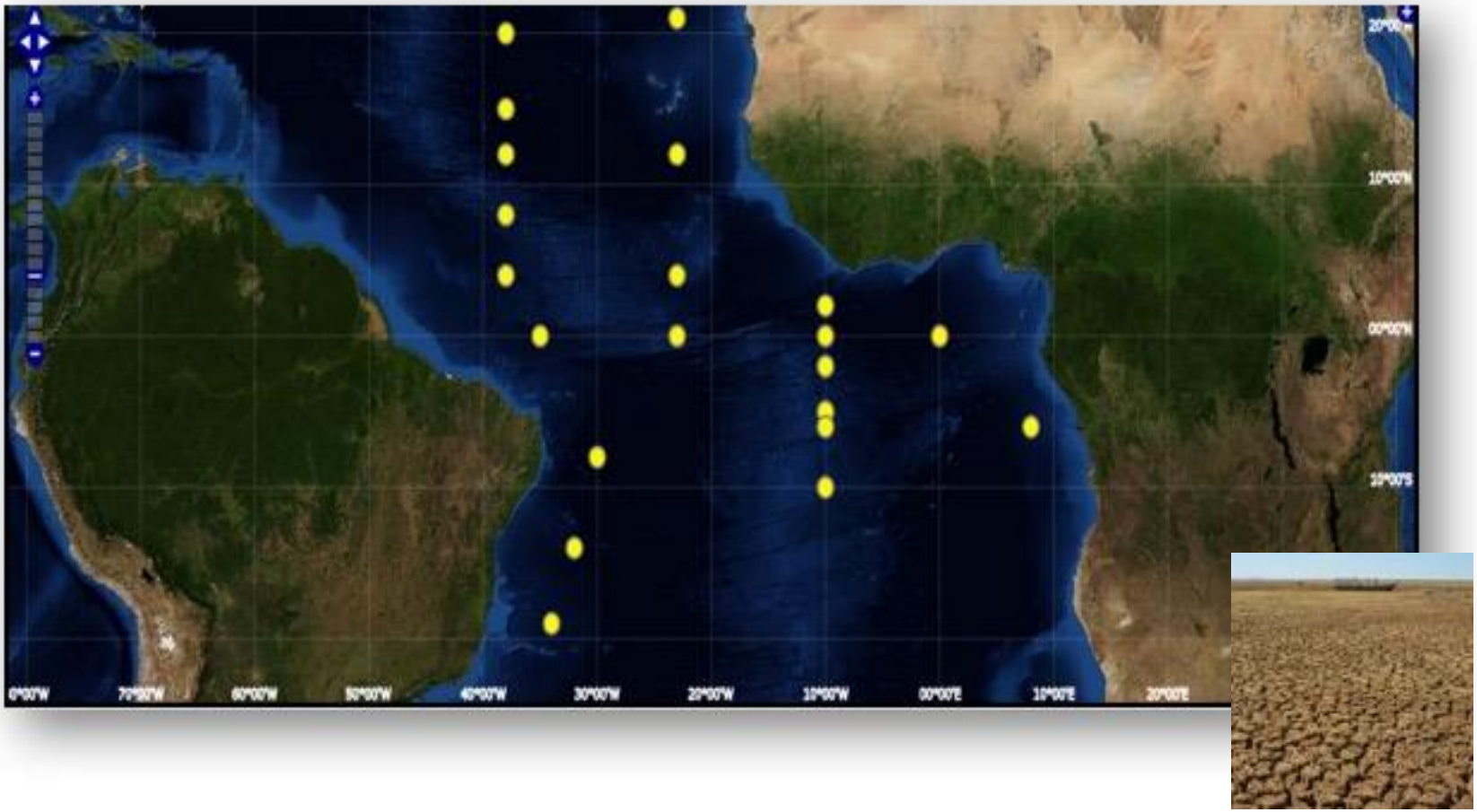




Predição de séries temporais usando aprendizado de máquina



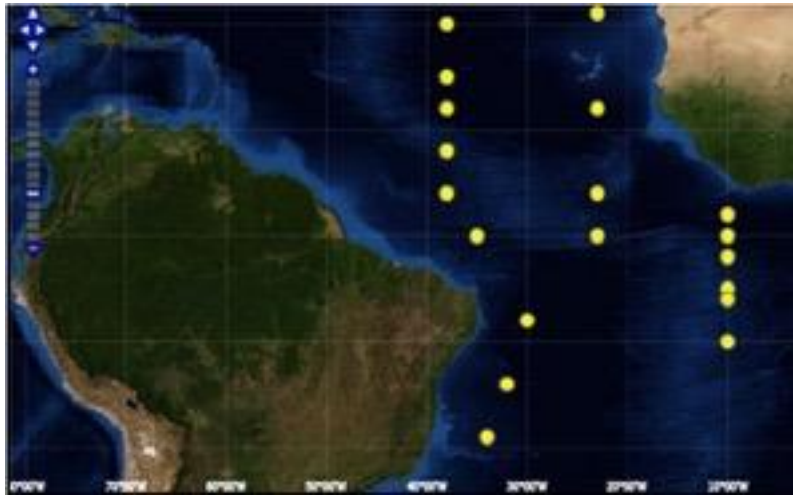
Prediction of sea surface temperature in South Atlantic Ocean

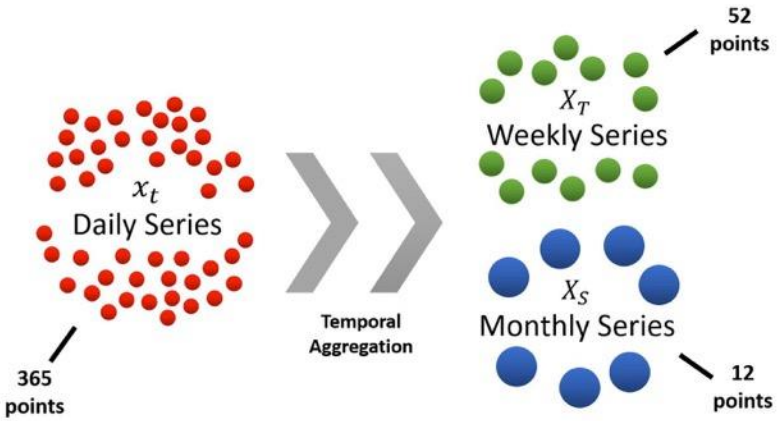


Spatial-time prediction

Time-Series Prediction

- Long term prediction of sea surface temperature





Time-Series Prediction – Results

- Effect of temporal aggregation for long-term prediction of sea surface temperature

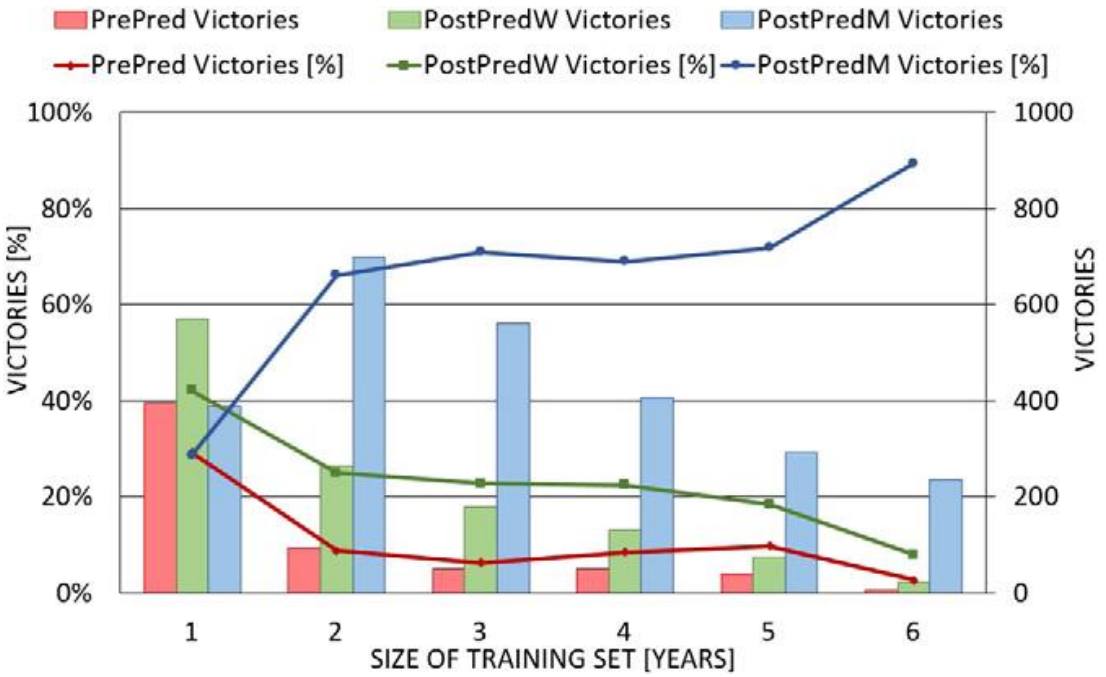


Fig. 8. Graphic of the victories of each prediction approach regarding their performances in generating up to twelve monthly aggregated forecasts.

Rebecca Salles
Scientific initiation

Time-Series Prediction – Results

- Framework for analysis of prediction performance compared to linear models

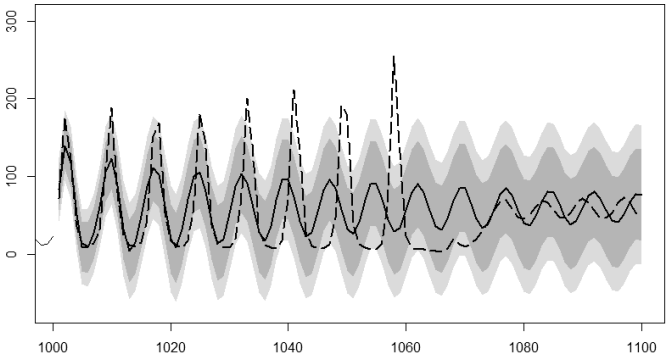


Fig. 2: ARMA predictions (solid line) for the time series A of the Santa Fe Competition. The actual time series values are represented by the dashed line.

TABLE III: Rankings of the top 25 results of the chosen competition datasets including results from TSPred R-package

Rank	Santa Fe				EUNITE		CATS			NN3		NN5	
	Dataset A		Dataset D							Dataset A		Dataset A	
	index	NMSE	index	NMSE[1]	Participant	MAPE [%]	Participant	E1	E2	Participant	Mean SMAPE	Participant	Mean SMAPE
1	W	0.02	ZH	0.08	Chih-Jen Lin	1.982	Sarkka*	408	346	Illies*	15.18%	Andrawis	20.40%
2	Sa	0.08	TSPred$_{(ARIMA)}$	0.54	Esp	2.149	Cai*	441	402	Adeodato*	16.17%	Vogel	20.50%
3	M	0.38	U	1.30	Brockmann	2.498	Kurogi*	502	418	Flores*	16.31%	D'yakonov	20.60%
4	L	0.45	TSPred$_{(PR)}$	1.61	TSPred$_{(PR)}$	2.779	Hu*	530	370	Chen*	16.55%	Rauch	21.70%
5	U	0.62	Z	4.80	Zivcak	2.873	Palacios-Gonzalez	577	395	D'yakonov	16.57%	Luna	21.80%
6	A	0.71	C	6.40	Kowalczyk	2.985	Maldonado*	644	542	Kamel*	16.92%	Wichard	22.10%
7	McL	0.77	W	7.10	Lewandowski	3.223	Simon*	653	351	Abou-Nasr	17.54%	Gao	22.30%
8	TSPred$_{(ARIMA)}$	0.90	S	17.00	Kowalczyk	3.264	Verdes*	660	442	Theodosiou*	17.55%	Puma-Villanueva	23.70%
9	TSPred$_{(PR)}$	0.99			Ortega	3.380	Chan*	676	677	TSPred$_{(ARIMA)}$	17.79%	Dang	25.30%
10	N	1.00			King	3.388	Wichard*	725	222	de Vos	18.24%	Pasero	25.30%
11	P	1.30			Lotfi	3.389	Beliaev*	928	762	Yan	18.58%	Adeodato	25.30%
12	Can	1.40			Guijarro	3.421	Kong	954	994	C49	18.72%	undisclosed	26.80%
13	K	1.50			Weizenegger	3.694	Wang	1037	402	Perfilieva*	18.81%	undisclosed	27.30%
14	Sw	1.50			TSPred$_{(ARIMA)}$	3.820	Cellier*	1050	278	Kurogi*	19.00%	TSPred$_{(ARIMA)}$	27.80%
15	Y	1.50			Boger	3.958	Crone*	1156	995	Beadle	19.14%	Tung	28.10%
16	Car	1.90			Bontempi	3.997	TSPred$_{(ARIMA)}$	1173	917	Lewicke	19.17%	undisclosed	33.10%
17					Pelikan	4.348	Acernese*	1247	1229	Sorjamaa*	19.60%	undisclosed	36.30%
18					Brockmann	4.373	Yen-Ping*	1425	894	Isa	20.00%	undisclosed	41.30%
19					Pelikan	4.437	TSPred$_{(PR)}$	7387	6778	C28	20.54%	TSPred$_{(PR)}$	41.50%
20					Rivieccio	4.502				Duclos-Gosselin	20.85%	undisclosed	45.40%
21					Brockmann	4.580				Papadaki*	22.70%	undisclosed	53.50%
22					Ivakhnenko	4.653				Hazarika	23.72%		
23					Brockmann	4.712				C17	24.09%		
24					Brockmann	5.087				Njimi*	24.90%		
25					Brockmann	5.425				Pucheta*	25.13%		

* et al.
[1] NMSE error for the 15 first predicted observations

Aspectos espaço-temporais

Distribuição varia no espaço e tempo

Ipeadata - PIB Estadual a preços constantes
Brasil - 2006 - Malha: Estados

Ipeadata - PIB Estadual a preços constantes
Brasil - 2010 - Malha: Estados

>= 277.027.984
< 277.027.984
< 238.424.656
< 199.821.328
< 161.218.001
< 122.614.673
< 84.011.345
< 45.408.017
< 6.804.690

Omisso

Mudança na faixa de valores

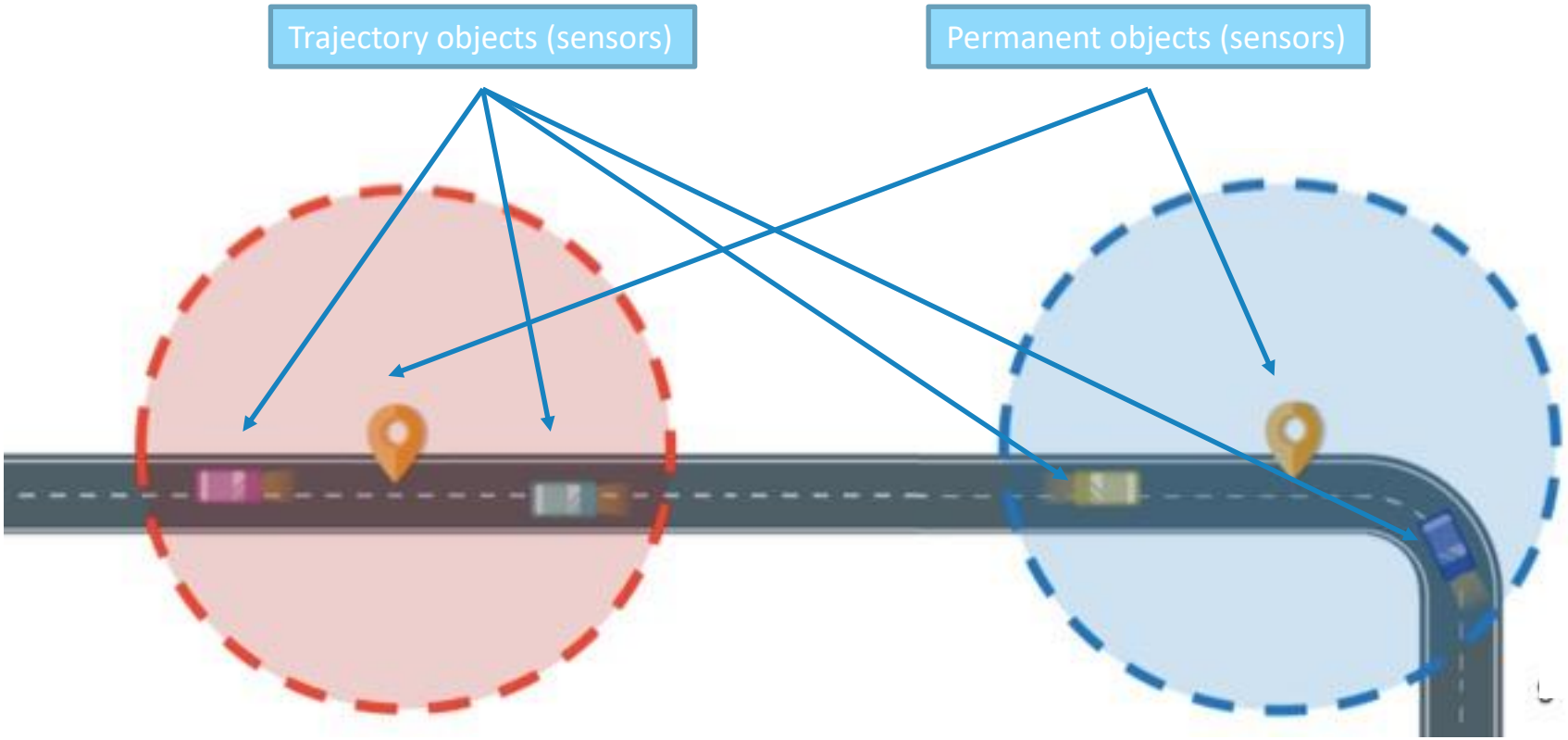
Sigla	Estado	2006	2010
AC	Acre	6.633.867	8.476.515
AL	Alagoas	20.802.615	24.574.808
AM	Amazonas	50.816.007	59.779.292
AP	Amapá	6.804.690	8.265.965
BA	Bahia	131.479.024	154.340.458
CE	Ceará	64.306.577	77.865.415
DF	Distrito Federal	125.765.530	149.906.319
ES	Espírito Santo	66.563.030	82.121.834
GO	Goiás	78.044.303	97.575.930
MA	Maranhão	37.195.271	45.255.942
MG	Minas Gerais	302.431.433	351.380.905
MS	Mato Grosso do Sul	34.311.309	43.514.207
MT	Mato Grosso	46.453.960	59.599.990
PA	Pará	69.415.228	77.847.597
PB	Paraíba	26.429.318	31.947.059
PE	Pernambuco	77.462.625	95.186.714
PI	Piauí	17.958.480	22.060.161
PR	Paraná	179.844.892	217.289.677
RJ	Rio de Janeiro	354.234.639	407.122.794
RN	Rio Grande do Norte	28.261.660	32.338.895
RO	Rondônia	17.978.571	23.560.644
RR	Roraima	5.007.053	6.340.601
RS	Rio Grande do Sul	217.001.407	252.482.597
SC	Santa Catarina	132.634.660	152.482.338
SE	Sergipe	19.962.748	23.932.155
SP	São Paulo	1.024.208.900	1.247.595.927

Spatial-time series

Let $P = \{p_1, p_2, ..., p_m\}$ be a set of positions, a **spatial-time series** d is a couple (p, t) where $p \in P$ is a position and t is the associated time series.

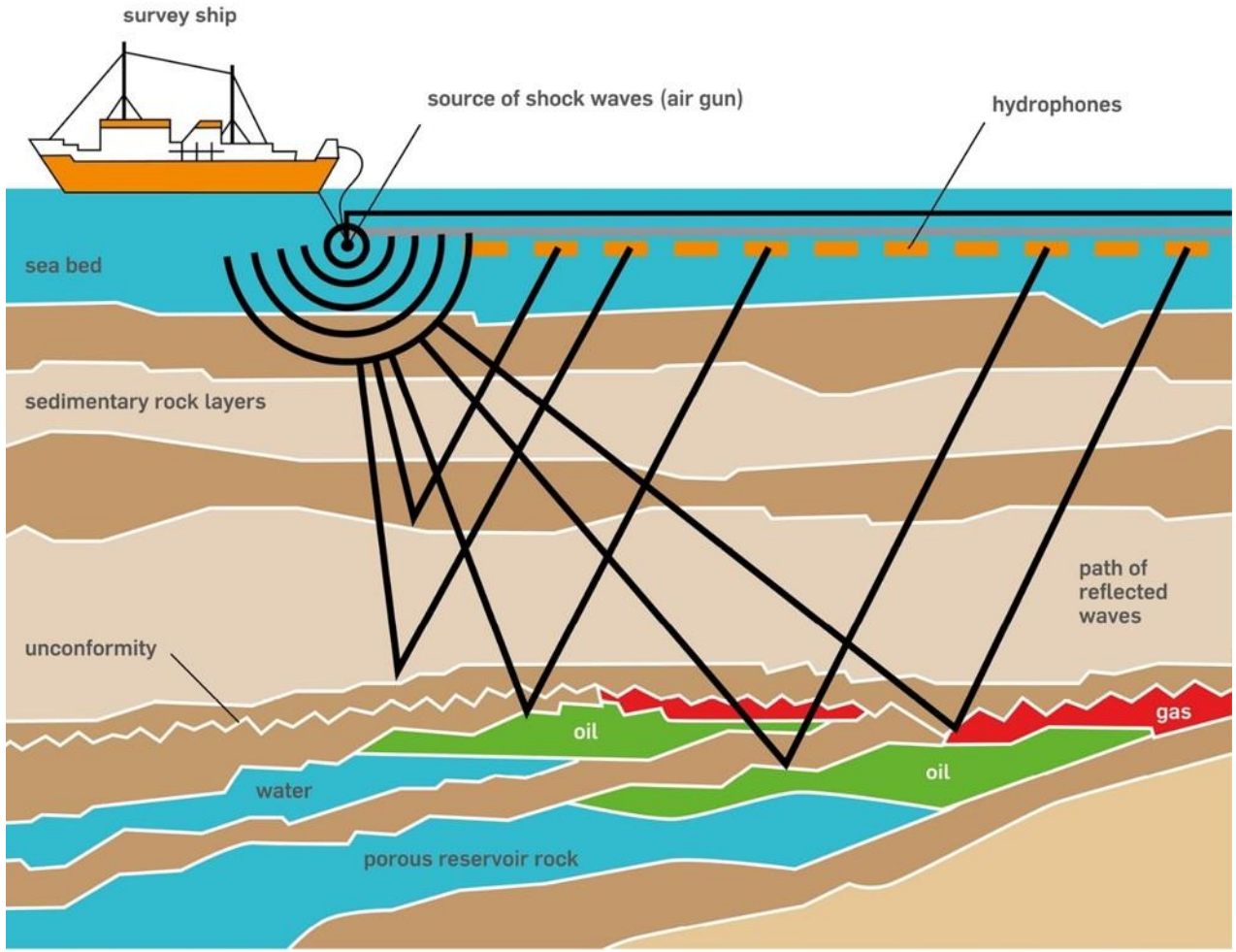
A **spatial-time series dataset** D is a set of spatial-time series $\{d_j\}$.

Given a $d = (p, t)$, if p varies according to time, d is a trajectory object, otherwise, d is a permanent object.



Análise na sísmica

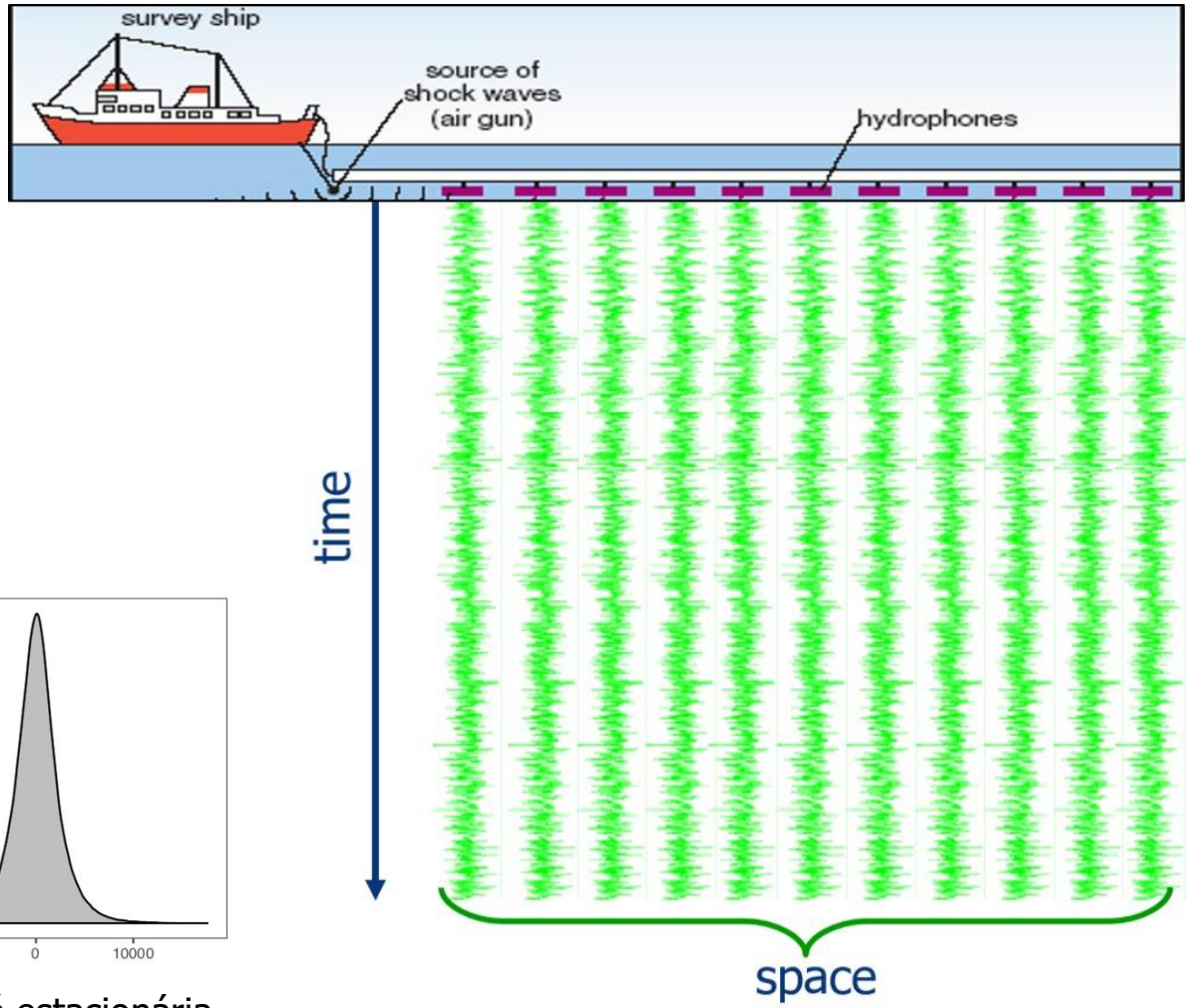
Séries espaço-temporais têm uma posição associadas a sensores

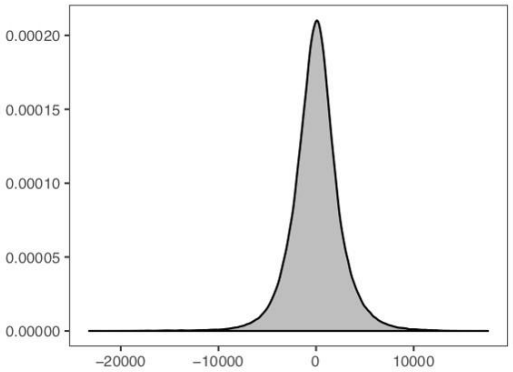


Source: https://krisenergy.com/company/about-oil-and-gas/exploration/

Análise na sísmica

Cada sensor está associada a uma série espaço-temporal

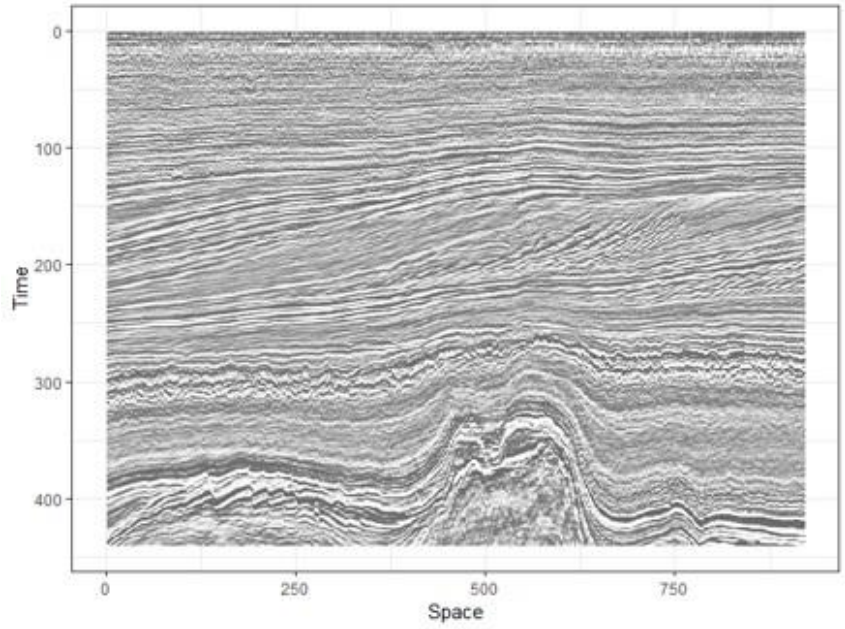




Cada série é estacionária

Não-estacionariedade no espaço-tempo

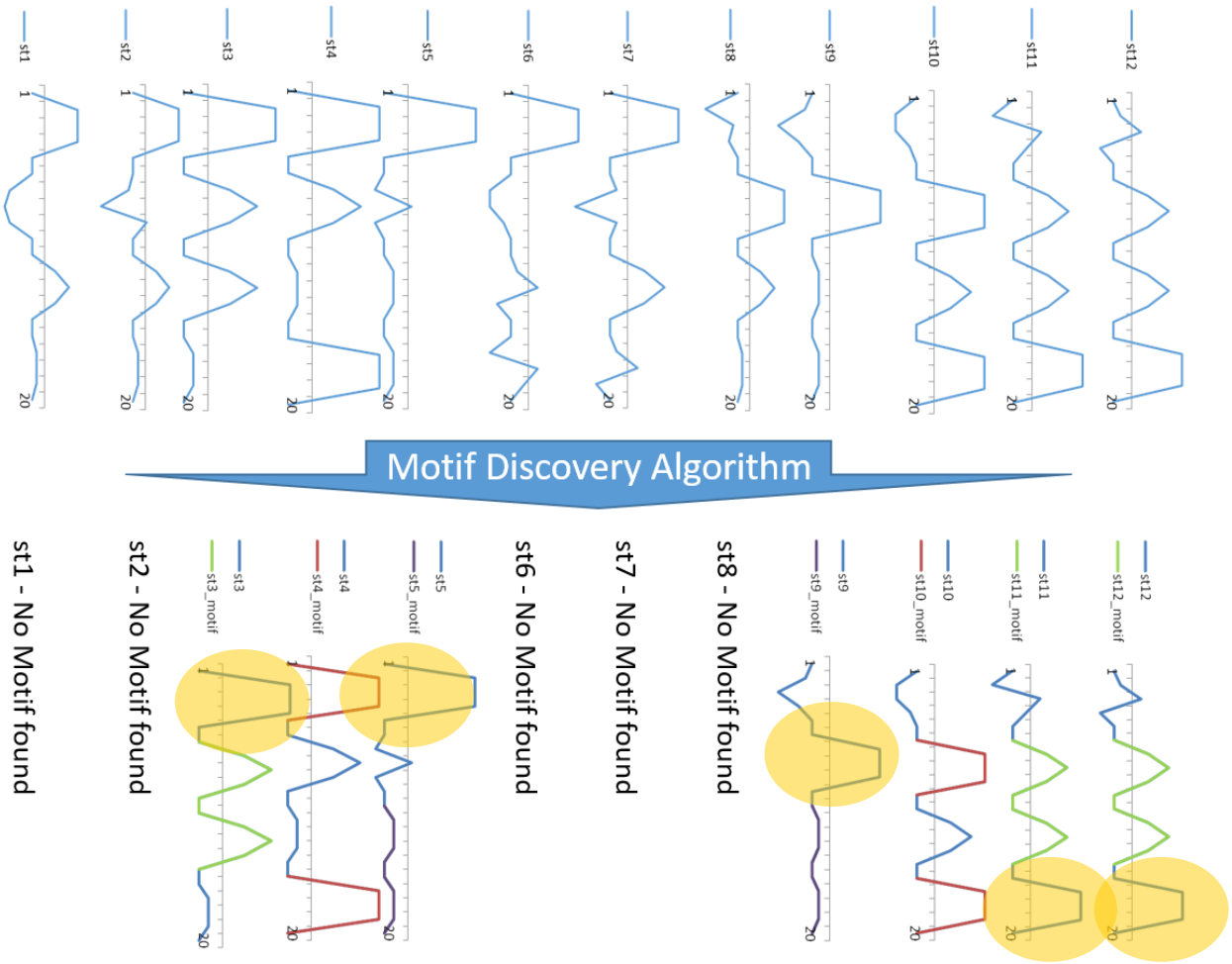
- Probabilidades diferentes no tempo-espaço
- Modelos especializados para regiões



[1] https://terranubis.com/datainfo/Netherlands-Offshore-F3-Block-Complete.

Discover motifs in spatial-time series

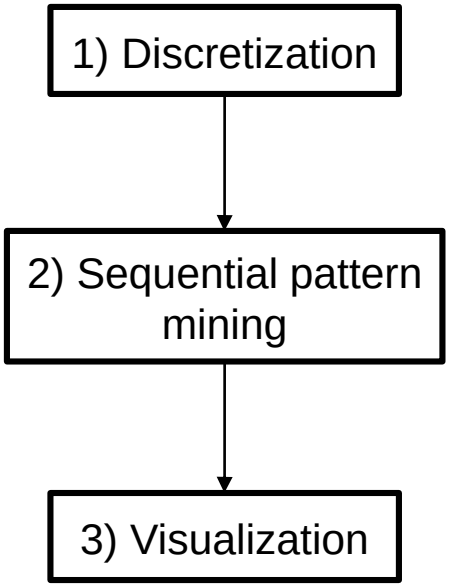
- Running motif discovery algorithm in single time series:
  - In some cases, no motif is found.
  - Similar shapes in the neighbors are not identified.



Traditional motif discovery algorithm applied in spatial-time series dataset. (i) red trapeziums and green triangles are identified motifs; (ii) blue trapeziums are not identified and not linked with red ones; (iii) blue triangles are not identified and not linked with green ones; (iv) purple shapes are not identified motifs

Approach 2: Sequence Mining

- Sequence pattern mining is used successfully to obtain insight from large volume of transactional databases.
- Scope of this work is the use of such technique to discover sequential patterns on seismic spatial-time series:
  - indexing technique used to discretize the input
  - adapted algorithm implemented to retrieve discovered patterns positions
  - results are presented over original seismic trace images to better evaluate the quality of results



A priori principle

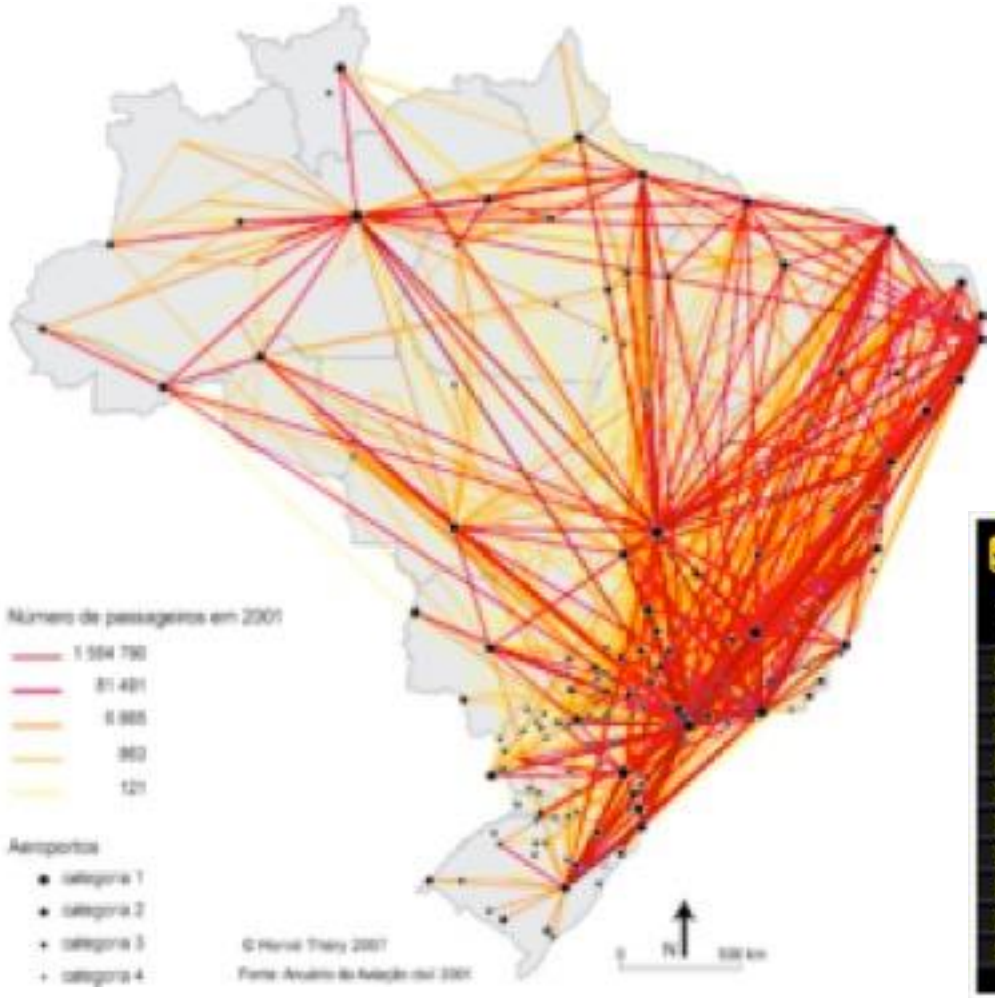
Time Square

Pattern Identification in Space-Time Series

D / t	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
v_1	**a**	b	c	d	t	q	**i**	g	**a**	h
v_2	k	l	m	n	p	q	**u**	s	t	v
v_3	w	**e**	**e**	x	y	m	**a**	r	d	a
v_4	h	**o**	**o**	g	**e**	i	e	**i**	c	b
v_5	**i**	j	k	l	**o**	z	n	**u**	z	p
v_6	**u**	**a**	r	s	t	∝	c	d	f	**a**

(sr_1) (sr_3) (sr_4) (sr_2)

(Business/Industrial)
Analysis of Flight Delays



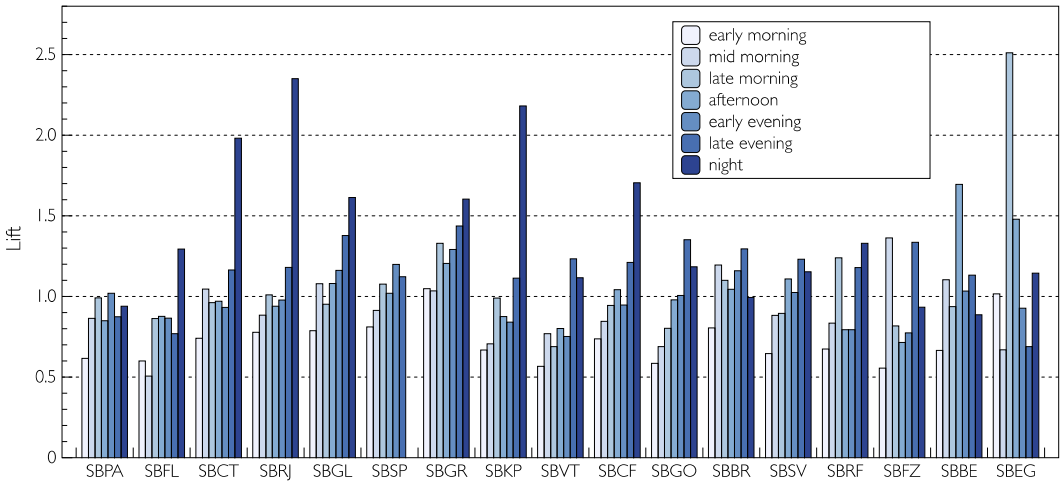
DEPARTURES

TIME	DESTINATION	FLIGHT	GATE	REMARKS
12:39	LONDON	CL 903	31	CANCELLED
12:57	SYDNEY	UQ5723	27	CANCELLED
13:08	TORONTO	IC5984	22	CANCELLED
13:21	TOKYO	AM 608	41	DELAYED
13:37	HONG KONG	IC5471	29	CANCELLED
13:48	MADRID	EK3941	30	DELAYED
14:19	BERLIN	AM5021	28	CANCELLED
14:35	NEW YORK	ON 997	11	CANCELLED
14:54	PARIS	MG5870	23	DELAYED
15:10	ROME	RI5324	43	CANCELLED

Analysis of delays in airports according to time

Flight Delays – Results

- Data warehouse
  - Brazilian National Flights
  - Meteorological condition
- Identification of frequent patterns that leads to delays



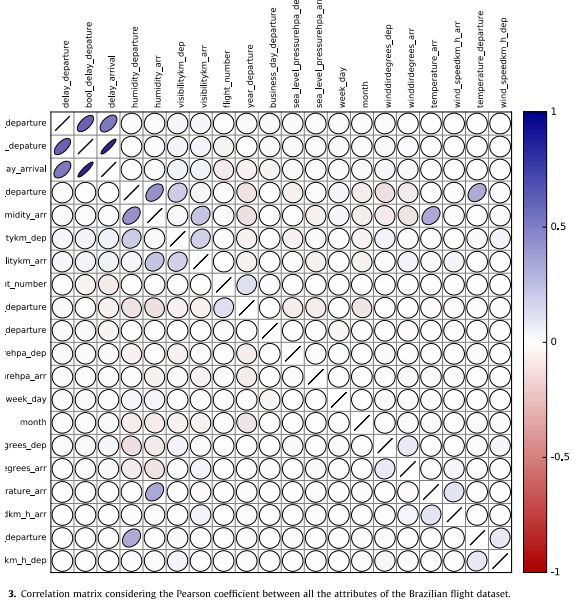
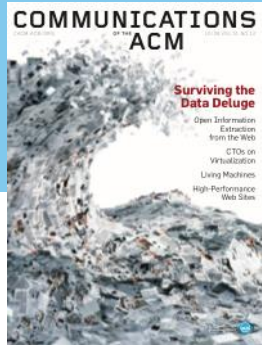


Fig. 7. Lift analysis of the rules containing the airport and the time of departure on the antecedent and a delay on the consequent – the airports are ordered from south to north.

3. Correlation matrix considering the Pearson coefficient between all the attributes of the Brazilian flight dataset.

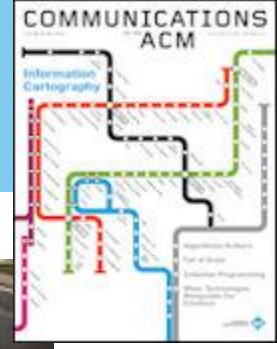
Alice Sternberg
master degree

Knowledge discovery in Time Series (domain)



- Big Data
  - Data deluge (volume and velocity)
  - Different data models (variability)
  - Science: astronomy, Seismic
  - Business/Persons: IoT, Flights
  - Government: Smart cities, Urban mobility
- Challenges for Knowledge Discovery
  - Data management
    - Data Preprocessing
    - Workflows
  - Data analysis
    - Prediction / Classification
    - Pattern Identification

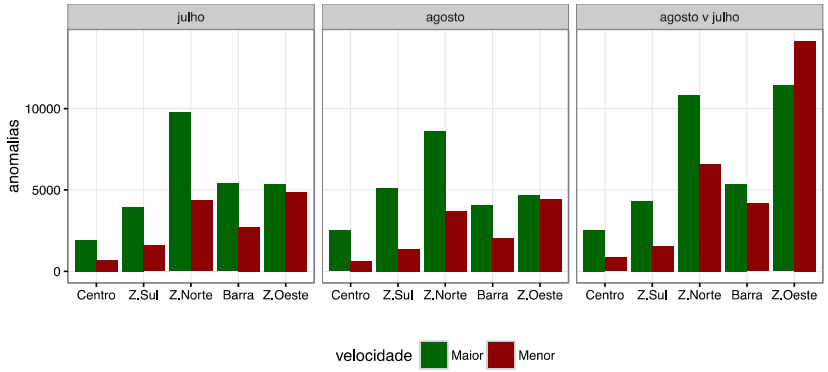
(Government)
Urban Mobility

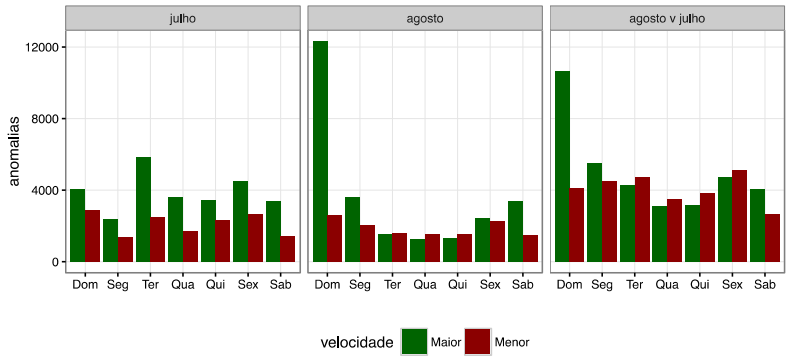


Buses as trajectory sensors: Analysis of Trajectory Data
Buses stops as permanent object sensors
(Spatial-time aggregation of buses data according to buses stops)

Urban Mobility – Results

- Data collection (done by UFF)
- Data Cleaning, Spatial-Time Aggregation
- Preliminary Analysis of Anomalies





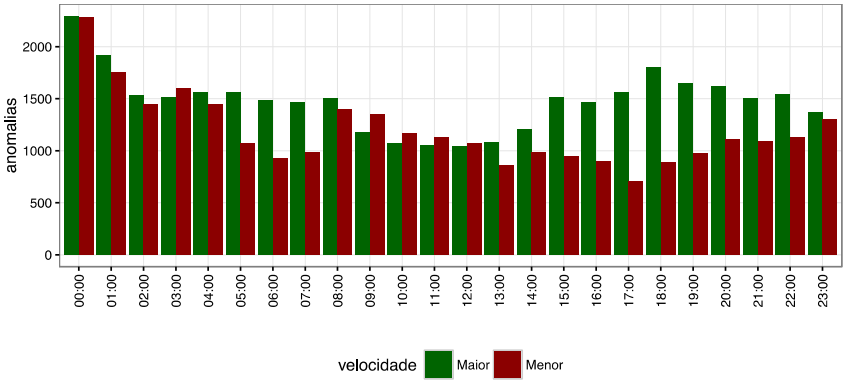


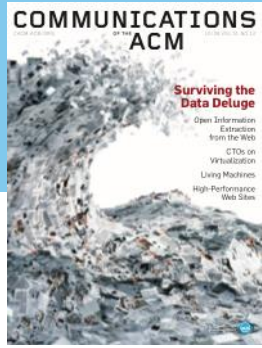
Figura 3. Anomalias identificadas por faixa de horário (ago v julho)

Ana Beatriz Cruz
Master degree

*Urban Mobility – **Research Opportunities***

- Persistence and Querying
- Trajectory or Aggregated analysis
- **Identification** of Patterns, Anomalies, and Drift

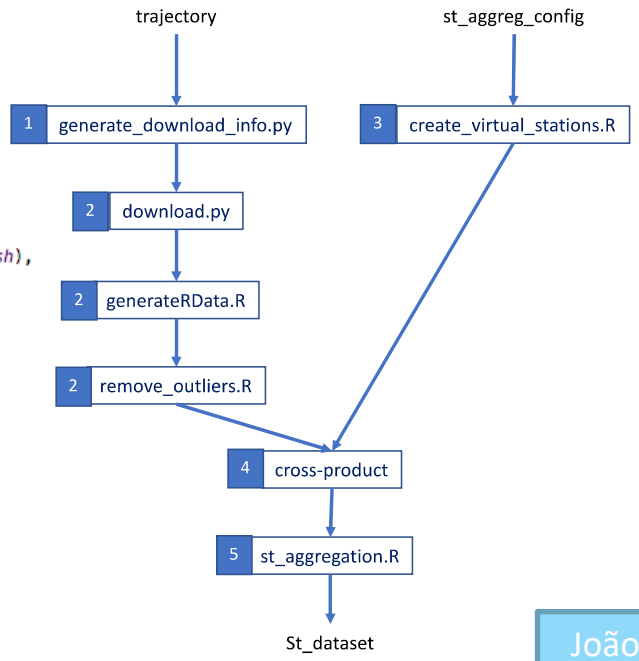
Knowledge Discovery in Time Series



- Big Data
  - Data deluge (volume and velocity)
  - Different data models (variability)
  - Science: astronomy, seismic
  - Business/Persons: IoT, flights
  - Government: smart cities, urban mobility
- Challenges for knowledge discovery
  - Data management
    - Data preprocessing
    - Workflows
  - Data analysis
    - Prediction / classification
    - Pattern identification

Parallel and Distributed Execution Using Spark

```
1  val trajectory: Relation = Relation(Schema(key, initialTime, endTime),
2    Tuple("copa-do-mundo-2014", "2014-06-01", "2014-07-31"))
3  val st_aggreg_config: Relation = Relation(Schema(radius, interval, busesMesh),
4    Tuple("10", "10", "malha-2014.csv"))
5  w = Workflow("2014CupAggregation", () => {
6    r1 = SplitMap(Activity("generate_download_info.py"), key, trajectory)
7    r2 = Map(Activity("download.py"), r1)
8    r3 = Map(Activity("generateRdata.R"), r2)
9    r4 = Map(Activity("remove_outliers.R"), r3)
10   r5 = Map(Activity("create_virtual_stations.R"), st_aggreg_config)
11   r6 = Query(CrossProduct, r4, r5)
12   result = Map(Activity("st_aggregation.R"), r6)
13 })
14 w.execute()
```



João Ferreira
Master degree

(a) (b)

Figura 2. Workflow para análise de tráfego durante a COPA de 2014 : a) Especificação do Workflow usando linguagem Scala; b) grafo mostrando as dependências entre as atividades

Research project in Management and Analysis of Spatial-Time Series

- Novel algorithms for prediction/classification and pattern identification
  - Motif identification
  - Tight spatial-time sequence mining
- Explore spatial-time series applications
  - Frequent pattern mining, Classification/Prediction
- Explore data management and parallel processing for mining non-stationary time/spatial-time series
  - Algebraic-based workflows for spatial-time series data mining using Spark

CEFET/RJ Team
(24 active students)

Ref	Student	Theme	Level
HB	Heraldo Pimenta Borges Filho	Discovering motifs restricted in space-time	D.Sc.
RC	Rocío Milagros Zorrilla Coz (coorientação)	Model selection for prediction of spatial-temporal datasets	D.Sc.
JF	Juan Humberto Leonardo Fabian (coorientação)	Machine learning recommendation of hpc parameters for multi-scale numeric simulations	D.Sc.
LC	Leonardo Mosqueira de Carvalho	Resilient approaches for concept drift in flight delay prediction	D.Sc.
RS	Rebecca Pontes Salles	Classification of event in streaming data	D.Sc.
PE	Paulo Augusto Neves de Carvalho Elias (coorientação)	Semi-automatic software merge using machine learning	D.Sc.
FC	Flavio Matias Damasceno de Carvalho	An approach for term weighting applied in sentiment analysis	D.Sc.
LB	Lais Ribeiro Baroni*	Discovering divergent association rules: a case study of malaria on legal amazon	D.Sc.
AA	Adalberto Mineiro de Andrade (coorientação)	Prediction of worldwide fertilizers consumption	M.Sc.
AF	Antonio Jose de Castro Filho	Sequence mining in 3d spatial-time series	M.Sc.
LE	Luciana Escobar Gonçalves Vignoli (coorientação)	A framework for event detection in time series	M.Sc.
CT	Claudio Marcio da Silva Teixeira	Particionamento horizontal para mineração de padrões frequentes em timetables	M.Sc.
LG	Lucas Giusti Tavares (coorientação)	An extensive analysis on flight delay prediction with concept drift	M.Sc.
RD	Raphael Dantas de Oliveira Pereira	Patent mining	M.Sc.
DL	Diego Silva de Salles	Detecção de eventos usando aprendizado de máquina	M.Sc.
AG	Arthur Ronald Ferreira Diogenes Garcia	ARAIMA: autoregressive adaptive integrated moving average	M.Sc.
FM	Flávio Pinheiro Marques (coorientação)	Using provenance to evaluate learning in educational games	M.Sc.
BP	Balthazar da Silva Cunha Paixão	Detecção de eventos nas notificações de síndrome respiratória aguda grave	TIC
DL-MS	Diego Carlos Lima & Matheus Soutto	ETL em patentes de tecnologia verde	TPF
LL-MM	Lucas Lima & Matheus Mencalha	A influência das campanhas pró-aleitamento materno na mortalidade neonatal	TPF
VV	Victor Valladares	Mineração de padrões frequentes sobre dados do enem	TPF
LM	Lyago Monteiro	Mineração de padrões frequentes sobre dados do hanafuda	TPF

Area

1 - Predictive Analytics

2 - Pattern Mining

3 - Data Analysis Workflows

4 - Data Science in Education

CEFET/RJ Team

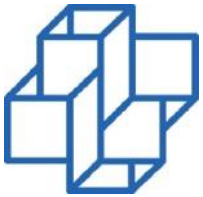
Dec/2016

Aug/2017

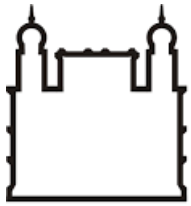
Dec/2018

Dec/2019

Main collaborators



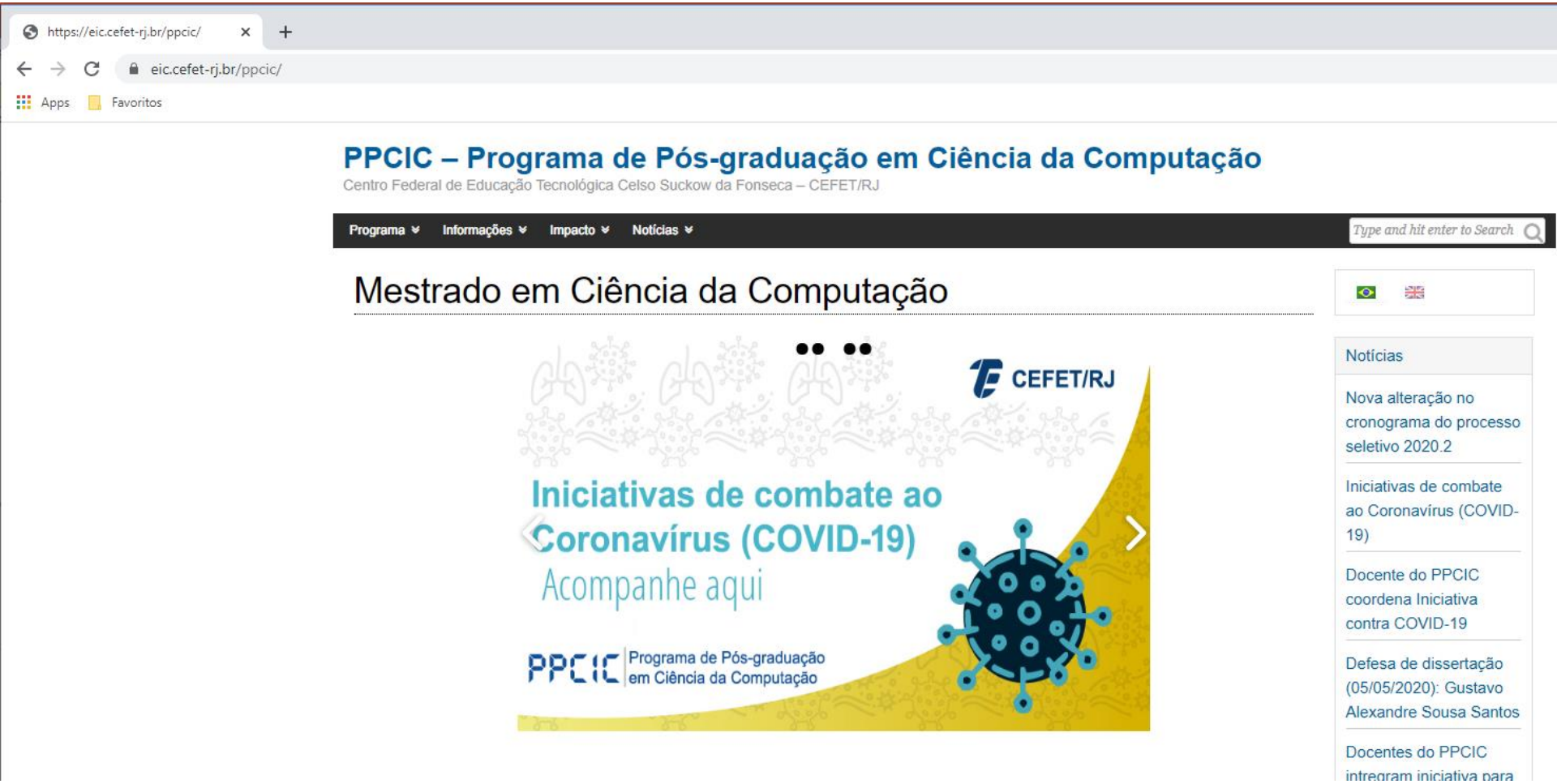
LNCC



Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

CEFET/RJ



Inscrições abertas para o mestrado em Ciência da Computação: http://eic.cefet-rj.br/ppcic/selecao
Inscrições abertas para o doutorado em Engenharia de Produção e Sistemas: http://pppro.cefet-rj.br

Recent Published Papers Related to the Project

- Cruz A. et al., 2017 - Detecção de anomalias no transporte rodoviário urbano. In SBBD.
- Ferreira J. et al, 2017 - Uma Proposta de Implementação de Álgebra de Workflows em Apache Spark no Apoio a Processos de Análise de Dados. In: BreSci
- Salles R. et al. 2017 - A Framework for Benchmarking Machine Learning Methods Using Linear Models for Univariate Time Series Prediction, IJCNN
- Marinho A. et al. 2017 - Deriving scientific workflows from algebraic experiment lines: A practical approach. Future Generation Computer Systems.
- Guedes G. et al. 2016 - Discovering top-k Non-Redundant Clusterings in Attributed Graphs. Neurocomputing.
- Sternberg A. et al., 2016 - An analysis of Brazilian flight delays based on frequent patterns. Transportation Research. Part E, Logistics and Transportation Review
- Salles R. et al, 2016 - Evaluating Temporal Aggregation for Predicting the Sea Surface Temperature of the Atlantic Ocean. Ecological Informatics.
- Machado E. et al, 2016 - Exploring machine learning methods for the Star/Galaxy Separation Problem. In: IJCNN
- Cruz A. et al, 2016 - Identificação de Motifs em Agregações de Séries Espaço-Temporais de Mobilidade Urbana. In: WTDBD/SBBD
- Campisano, R., Porto. F., Pacitti, E., Florent M., Ogasawara E., Spatial Sequential Pattern Mining for Seismic Data. In: SBBD
- Salles et al., 2015 - Evaluating Linear Models as a Baseline for Time Series Imputation. In: SBBD
- ...
- Ogasawara, E. et al., 2010 Adaptive Normalization: A Novel Data Normalization Approach for Non-Stationary Time Series. In: IJCNN.