Laboratório Nacional de Computação Científica

Time Series Event Detection

CEFET/RJ

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

Short Bio

- D.Sc. In Computer Science and Engineering at (COPPE/UFRJ) in 2011
- Professor at EIC - CEFET/RJ
  - Computer Science Department
  - Technical High-School in Computer Science
- Permanent Staff at
  - Postgraduate Program in Computer Science (PPCIC)
  - Postgraduate Program in Production Engineering and Systems (PPPRO)
- Member of IEEE, SBC, and ACM

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

Time series

- A time series is a sequence of observations of a phenomenon of interest collected over time

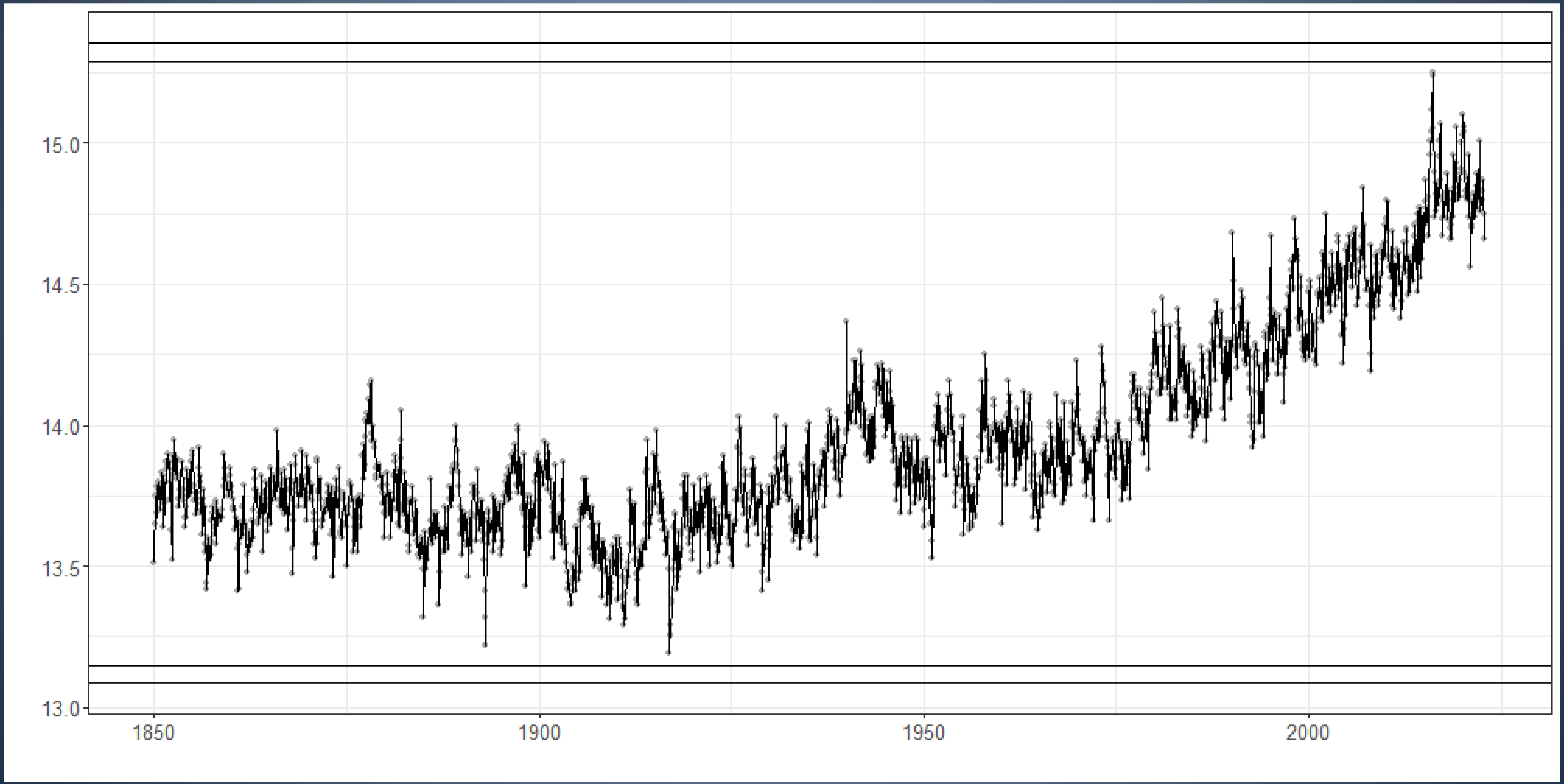
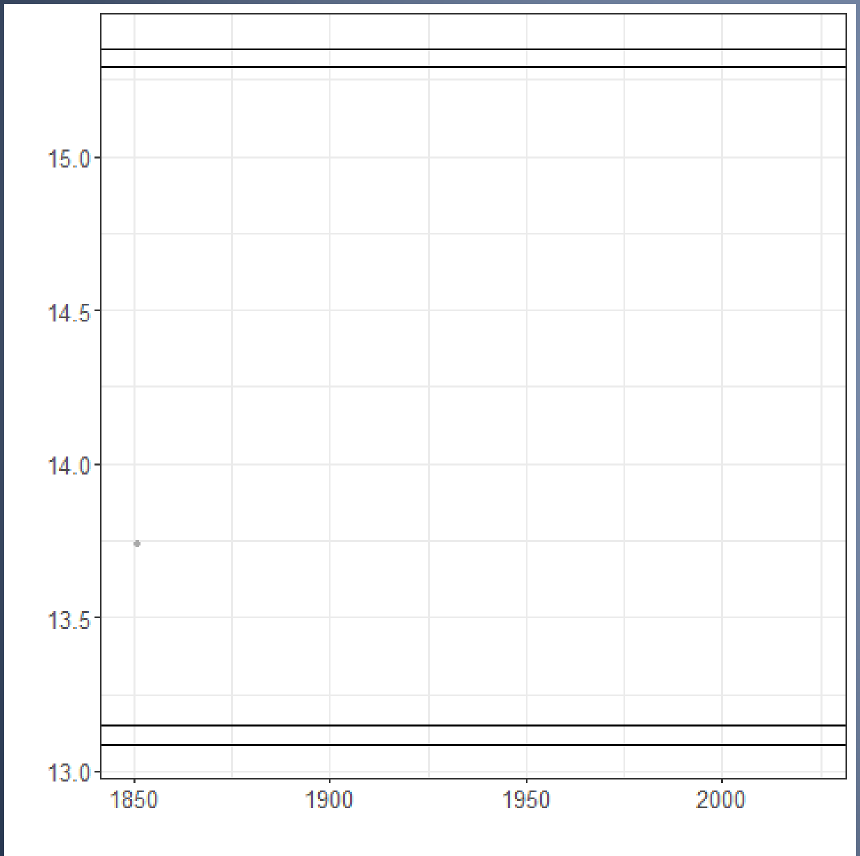
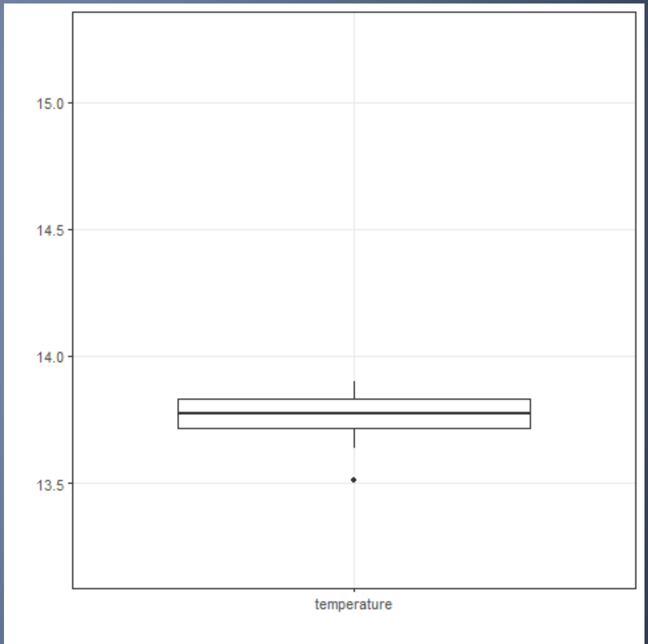


Image source: https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/global/time-series

Time series – online analysis

- Statistical properties may vary over time in streaming data



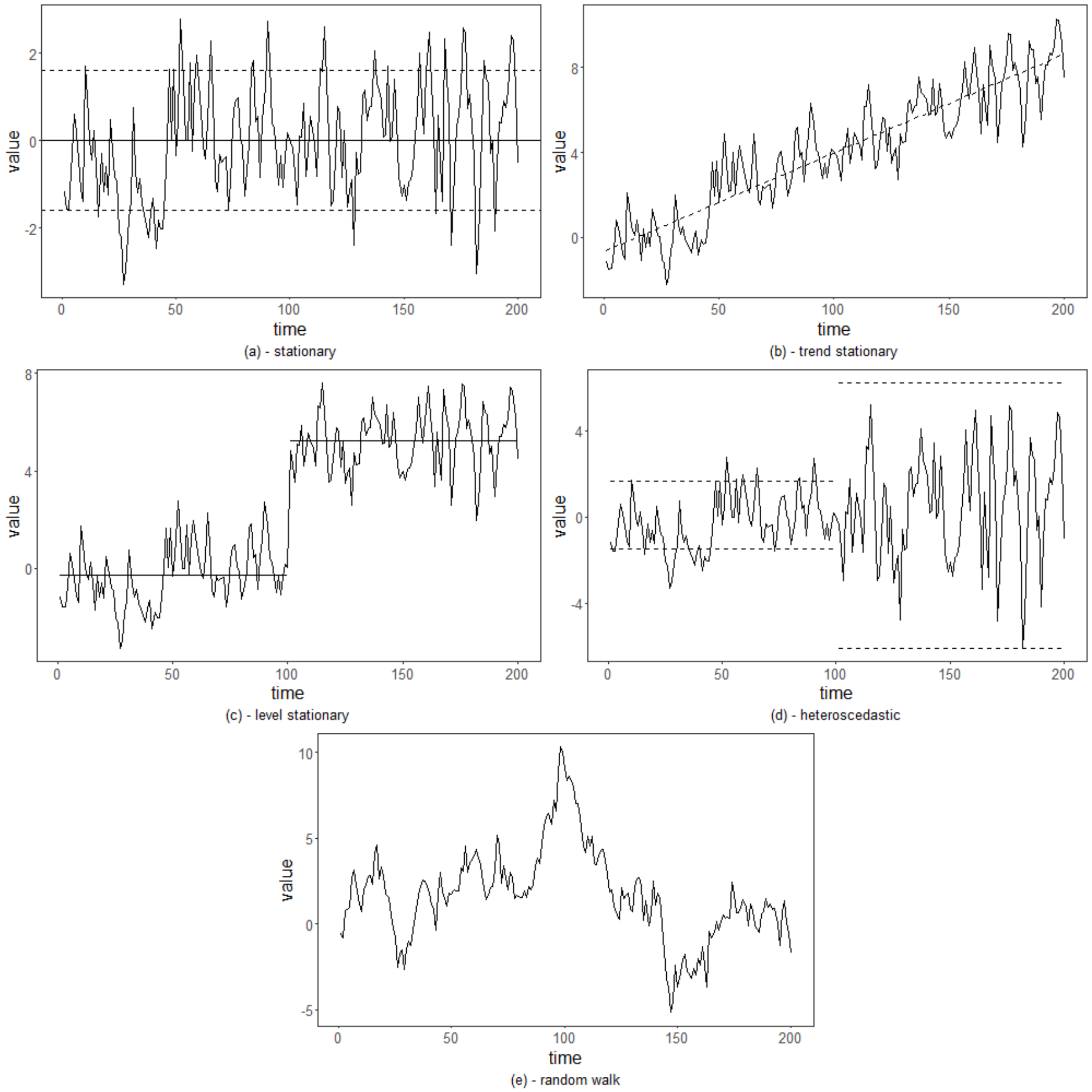


Stationarity

- Stationarity
  - Dataset D
  - Samples D_s from D
  - Statistical properties in D_s do not vary over time
    - mean, variance, covariance
- Non-stationarity
  - When stationary does not hold
- Data analytics methods
  - Most methods implicitly assume stationarity
- Pseudo-stationarity
  - When values of the time series are limited in a particular range during an interval

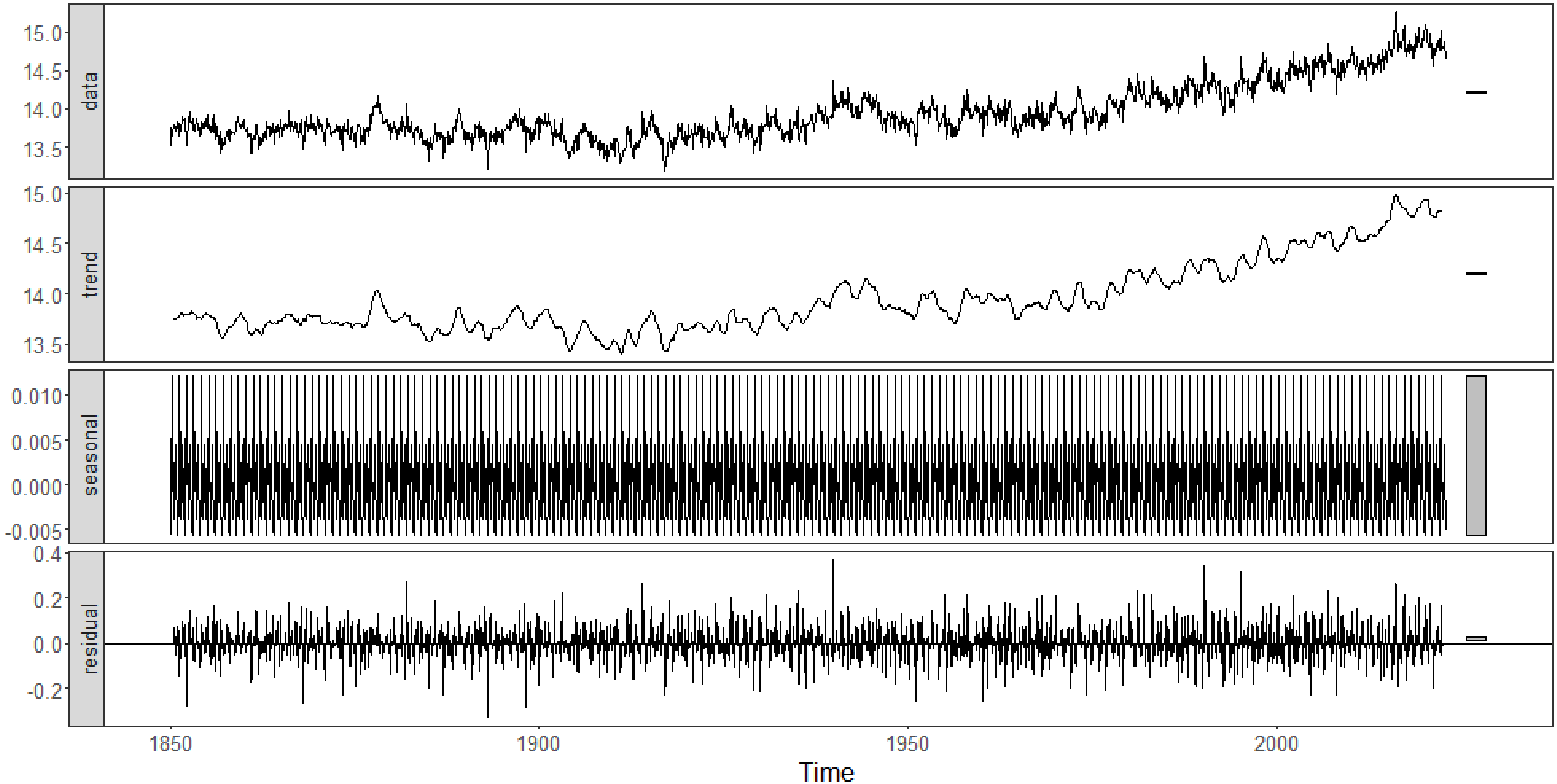
[1] R.H. Shumway e D.S. Stoffer, 2017, Time Series Analysis and Its Applications: With R Examples. 4 ed. New York, NY, Springer.

Stationarity and non-stationary time series



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.

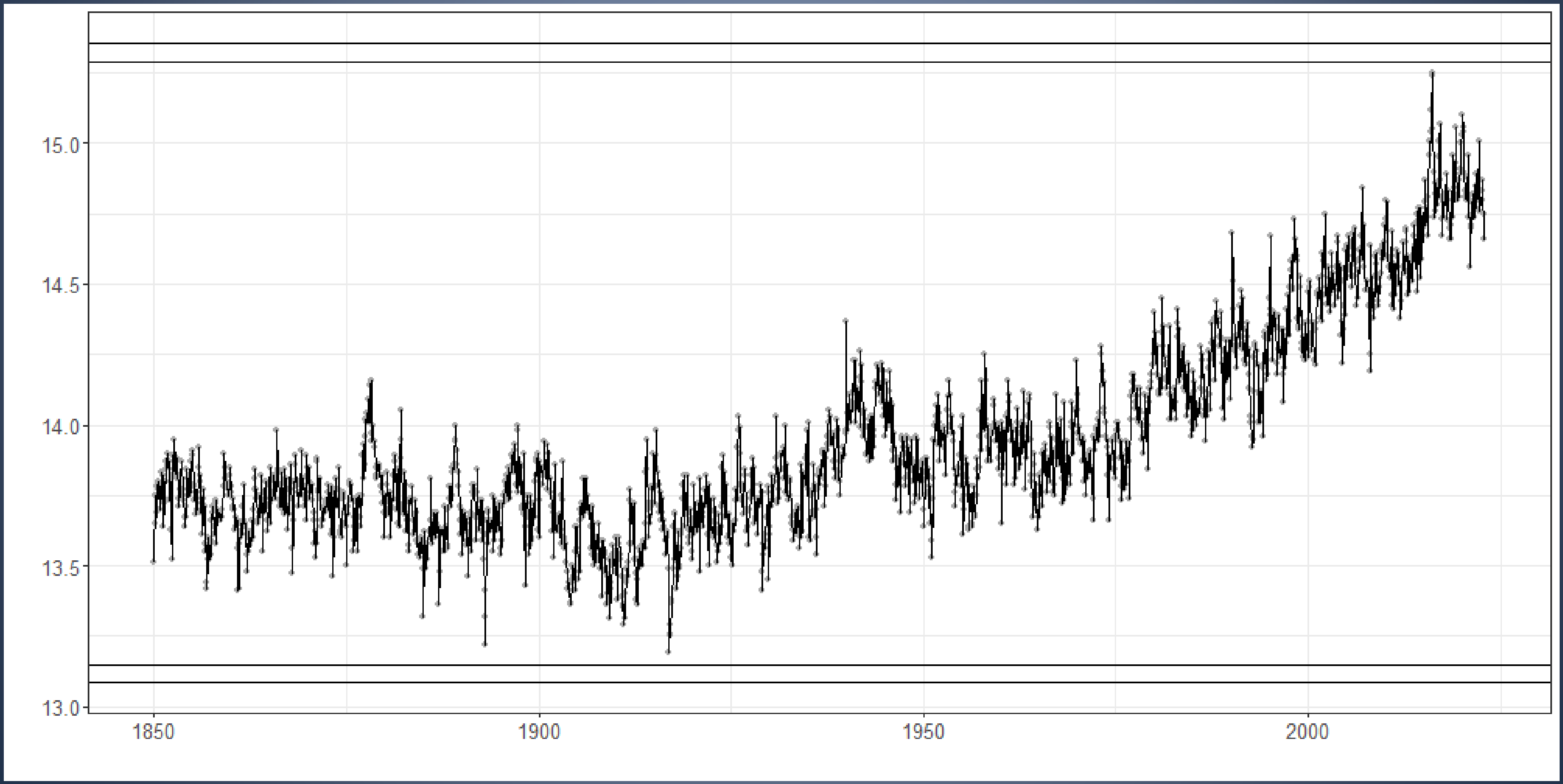
Time Series Components



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.

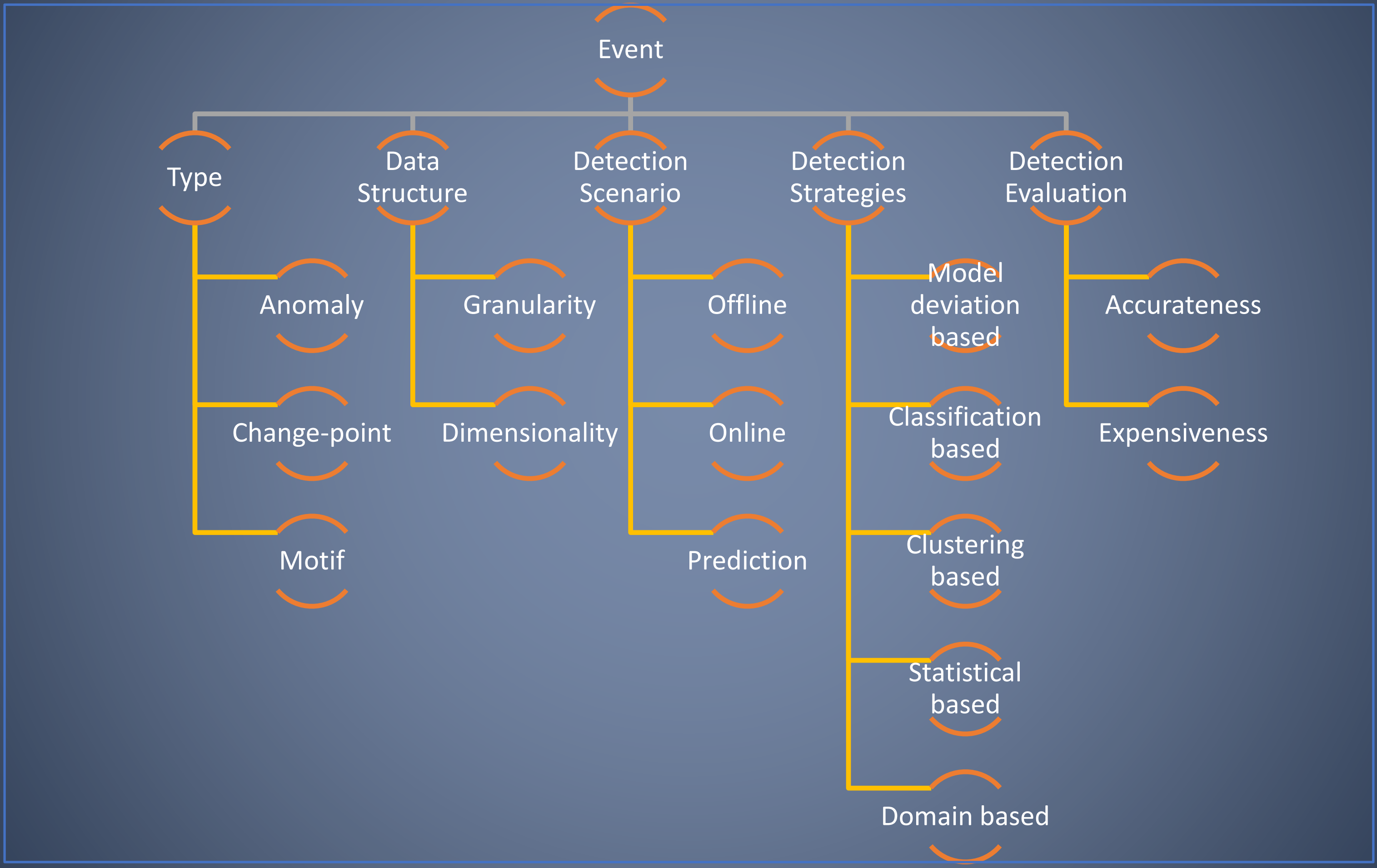
Events

- A point or an interval where a significant change in the time series behavior occurs
- Events may appear as anomalies, change points, or frequent patterns (motifs)



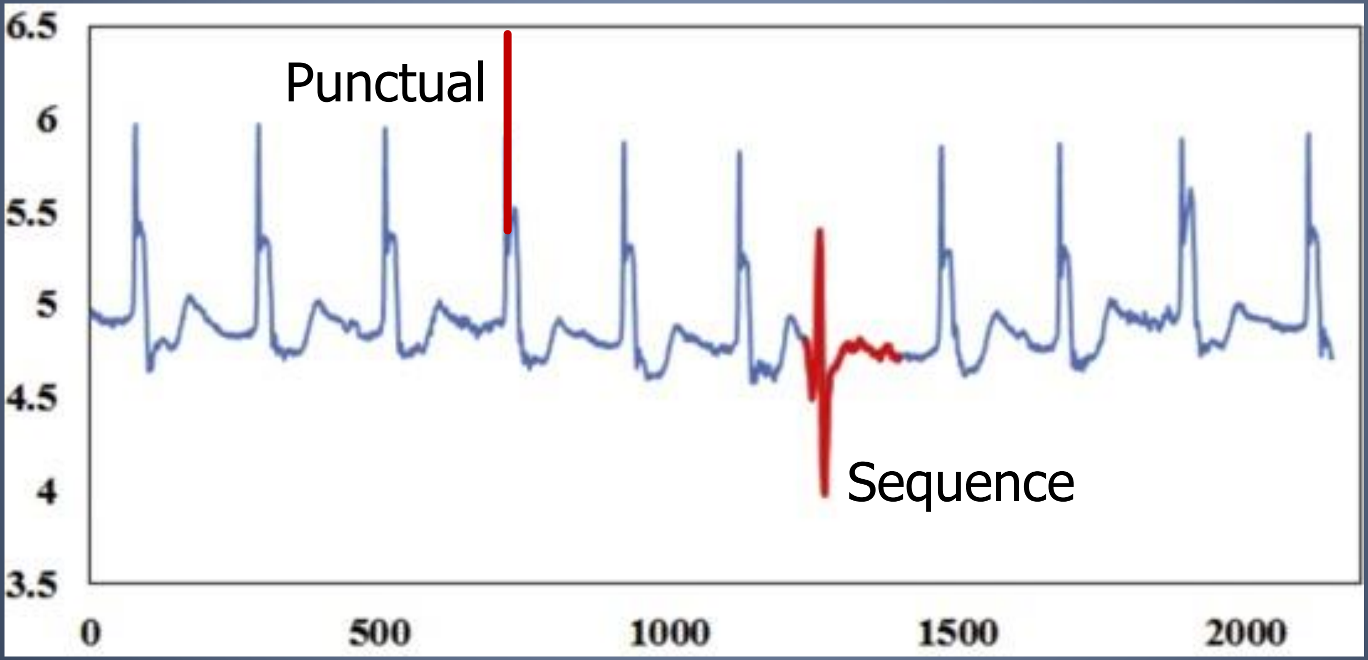
[1] V. Guralnik and J. Srivastava, 1999, Event Detection from Time Series Data, In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, p. 33–42
Image source: World Global Temperature, https://datahub.io/core/global-temp

Taxonomy of events



Anomalies

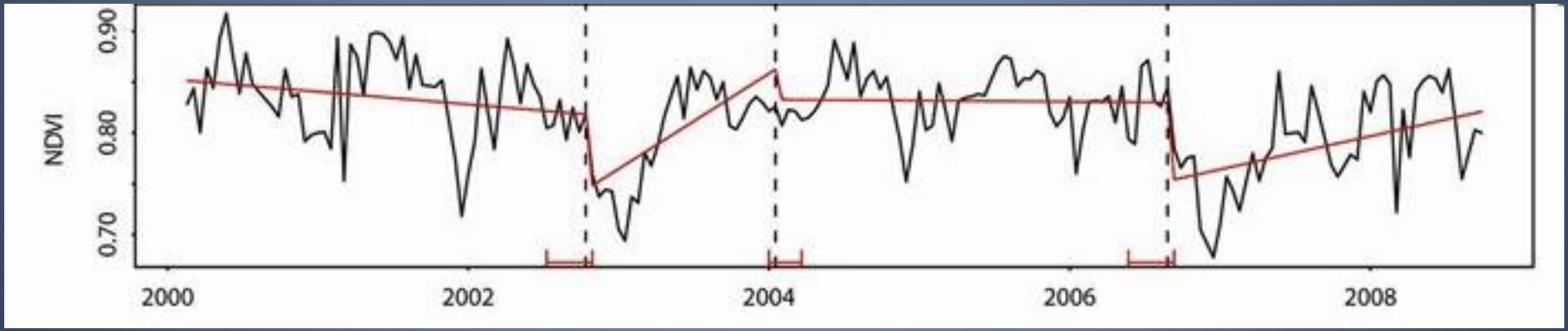
- A pattern or observation that does not conform to the expected behavior
- It can be categorized as punctual or interval (sequence)



[1]] V. Chandola, A. Banerjee, e V. Kumar, 2009, Anomaly detection: A survey, ACM Computing Surveys, v. 41, n. 3
(*) In this example, it can also be classified as a discord

Change Points

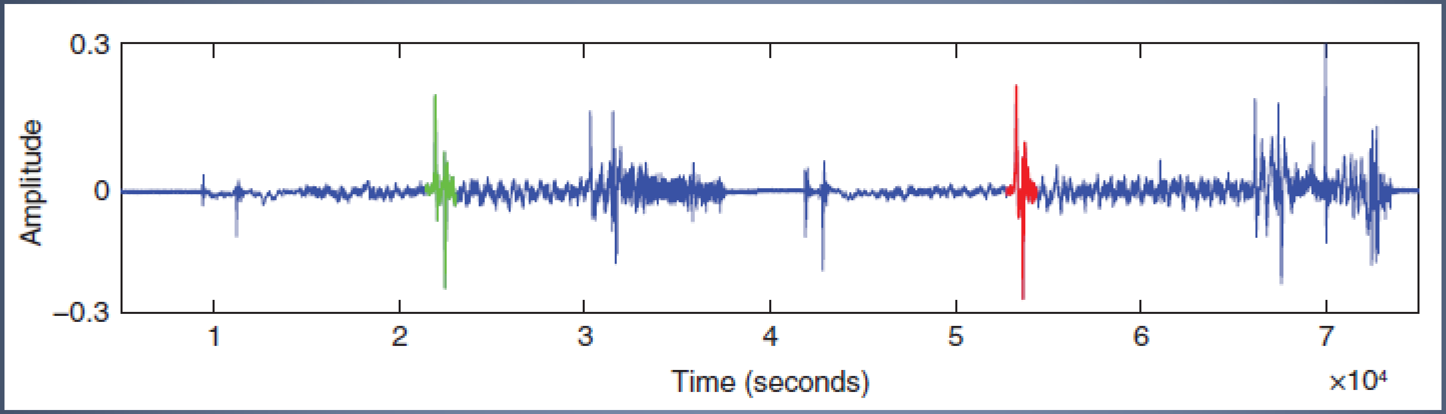
- Points (or time intervals) that mark significant change in time series behavior
- They separate different states in the process that generates the time series



[1] 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.
Image source: https://www.wur.nl/en/Research-Results/Chair-groups/Environmental-Sciences/Laboratory-of-Geo-information-Science-and-Remote-Sensing/Research/Integrated-land-monitoring/Change_detection_and_monitoring.htm

Motifs

- A pattern (unknown) that occurs a significant number of times in time series



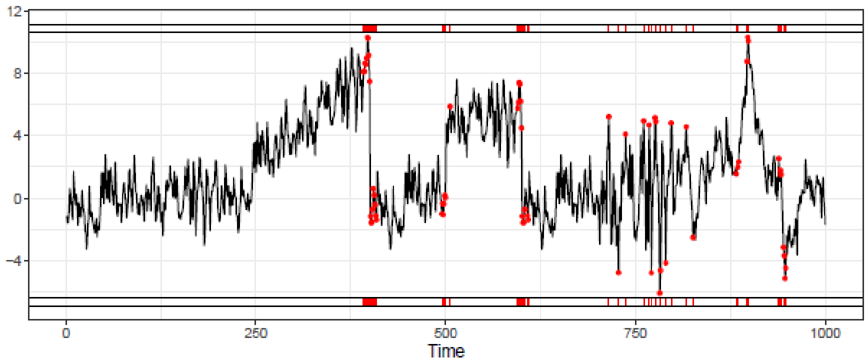
[1] P. Patel, E. Keogh, J. Lin, and S. Lonardi, "Mining motifs in massive time series databases," in Proceedings - IEEE International Conference on Data Mining, ICDM, 2002, pp. 370–377
[2] A. Mueen, "Time series motif discovery: Dimensions and applications," Wiley Interdisciplinary Reviews: Data Mining 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.

Summary of event detection initiatives

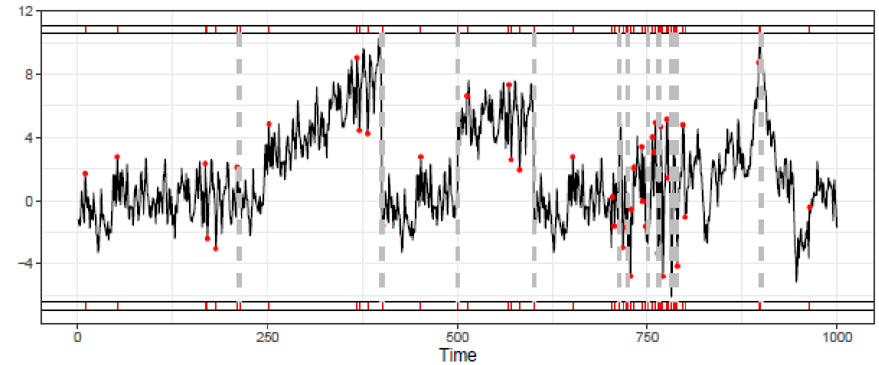
Anomaly detection	Finding unexpected behavior (deviations)
Change point detection	Finding change points It is related to finding drifts in time series
Motif detection	Identifying frequent patterns in time series

[1] Marco A.F. Pimentel, David A. Clifton, Lei Clifton, Lionel Tarassenko, A review of novelty detection, Signal Processing, Volume 99, 2014, Pages 215-249, ISSN 0165-1684, https://doi.org/10.1016/j.sigpro.2013.12.026.

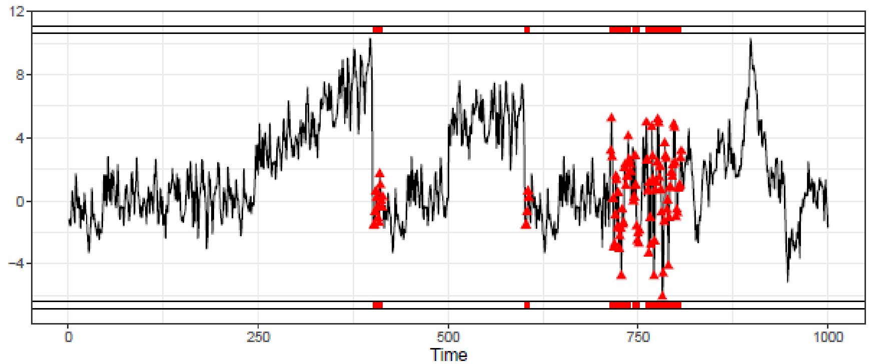
The many faces of event detection



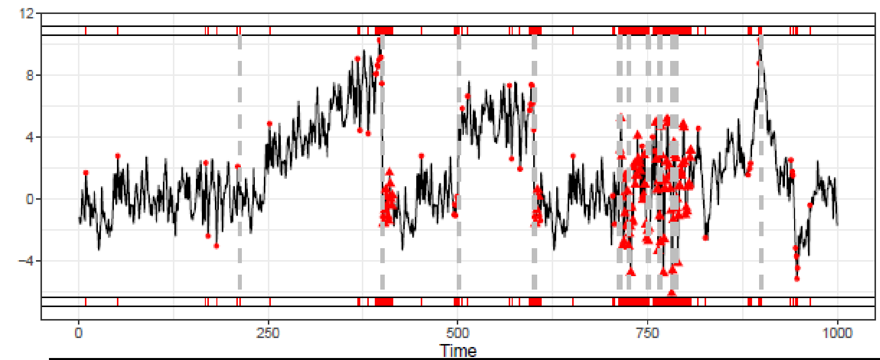
Method A: trend anomalies



Method B: trend anomalies & change points



Method C: volatility anomalies



Methods A,B & C:
trend anomalies, volatility anomalies and change points

[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)

Dimensionality

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)

Granularity

Monthly

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)

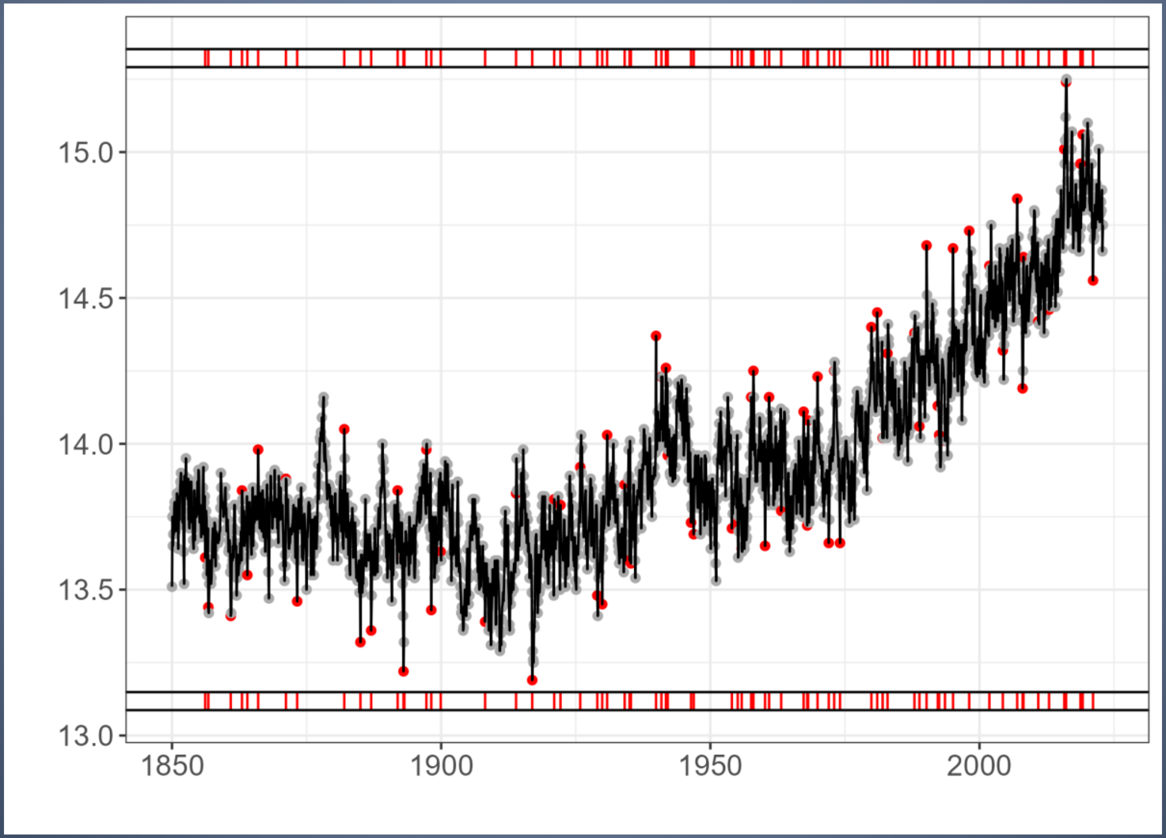
Yearly

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)

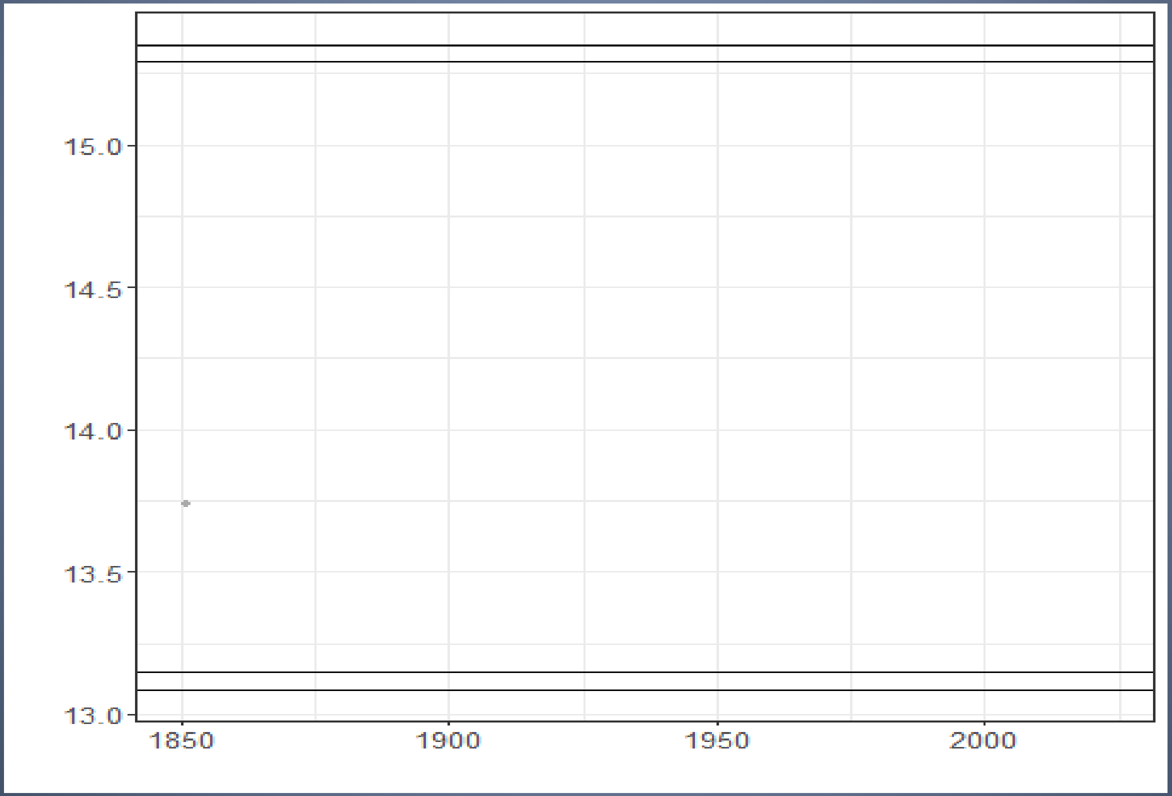
Scenarios - Offline

- Events are discovered after the time series has been collected
- It involves analyzing the time series retrospectively to identify patterns or changes that may indicate the occurrence of an event



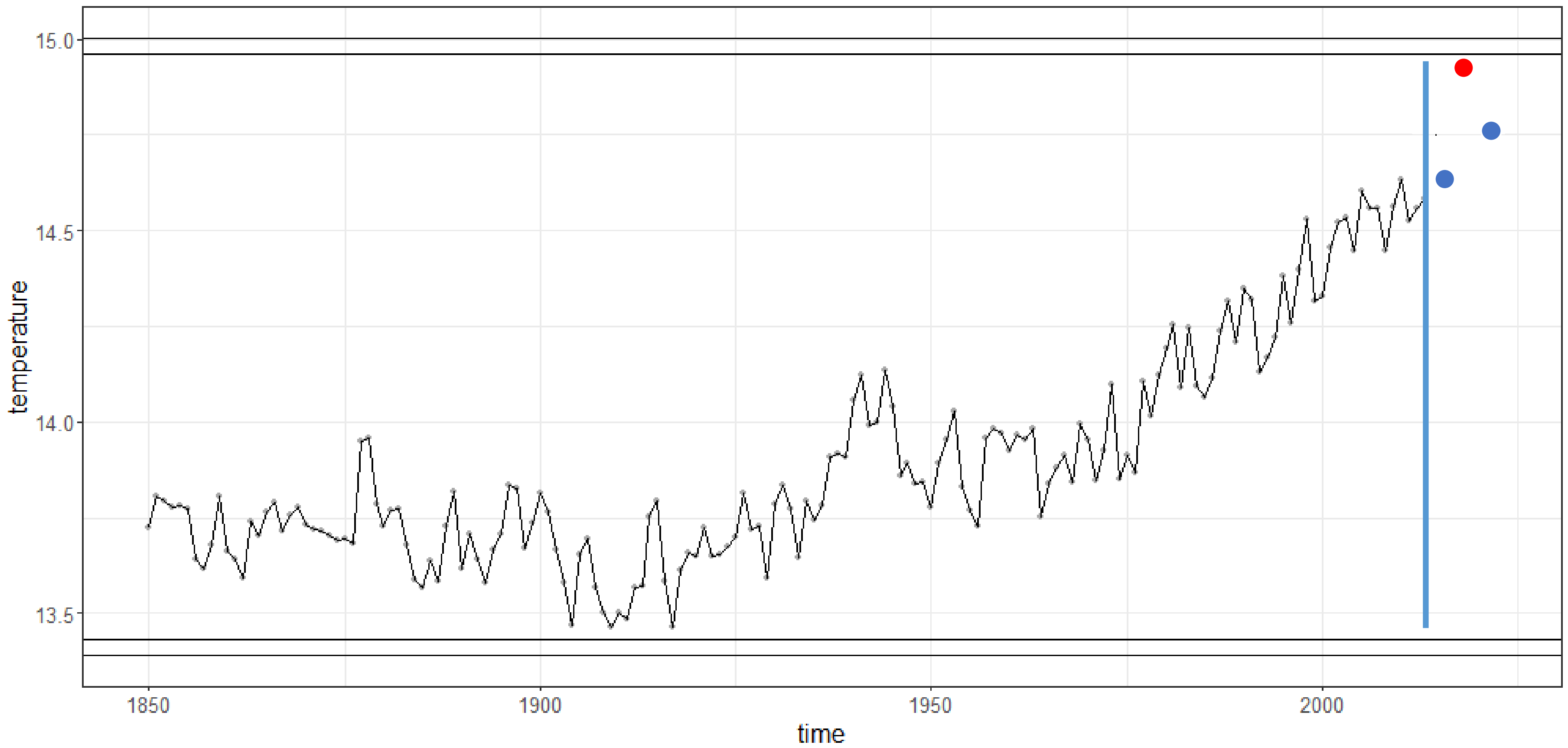
Scenarios - Online

- Events are discovered in a time series as they are collected
- It involves continuously monitoring the time series



Scenarios – Event Prediction

- At time t, predict that an event is going to occur at time $t + k$



(b)

Detection strategies

- Model deviation
- Classification-based
- Clustering-based
- Statistical-based
- Domain-based
- Ensemble

Model deviation

- Build a model (theory-driven or data-driven)
- Predict using model
- Analysis of differences

ω_t

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}

[1] V. Chandola, A. Banerjee, e V. Kumar, 2009, Anomaly detection: A survey, ACM Computing Surveys, v. 41, n. 3
[2] M. Gupta, J. Gao, C.C. Aggarwal, e J. Han, 2014, Outlier Detection for Temporal Data: A Survey, IEEE Transactions on Knowledge and Data Engineering, v. 26, n. 9, p. 2250–2267.
[1] R.A. Ariyaluran Habeeb, F. Nasaruddin, A. Gani, I.A. Targio Hashem, E. Ahmed, and M. Imran, 2019, Real-time big data processing for anomaly detection: A Survey, International Journal of Information Management, v. 45, p. 289–307.

Classification-based

- Labels: Supervised or semi-supervised learning

f_1

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}

Training

Testing

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}	$\hat{b}_{14}$

[1] G. Pang, C. Shen, L. Cao, and A.V.D. Hengel, 2021, Deep Learning for Anomaly Detection: A Review, *ACM Computing Surveys*, v. 54, n. 2
[2] A. Blázquez-García, A. Conde, U. Mori, and J.A. Lozano, 2021, A Review on Outlier/Anomaly Detection in Time Series Data, *ACM Computing Surveys*, v. 54, n. 3
[3] S. Thudumu, P. Branch, J. Jin, and J.J. Singh, 2020, A comprehensive survey of anomaly detection techniques for high dimensional big data, *Journal of Big Data*, v. 7, n. 1.

Clustering based

- Associate clusters to sequences
- Analyze differences with a representative sequence of a cluster

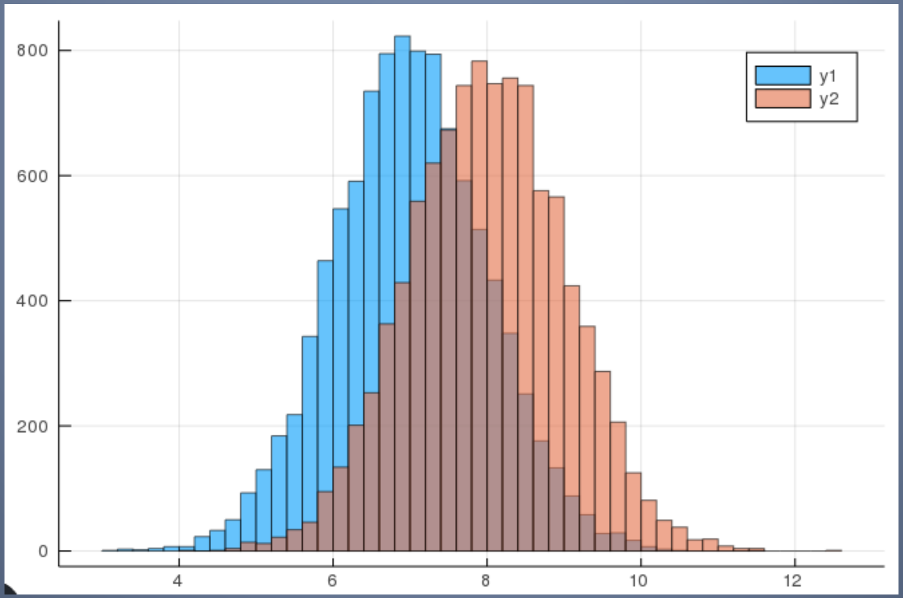
t	x_{t-4}	x_{t-3}	x_{t-2}	x_{t-1}	x_t	$\ddot{r}_c$	d_t
5	v_1	v_2	v_3	v_4	v_5	$\ddot{r}_1$	d_5
6	v_2	v_3	v_4	v_5	v_6	$\ddot{r}_1$	d_6
7	v_3	v_4	v_5	v_6	v_7	$\ddot{r}_1$	d_7
8	v_4	v_5	v_6	v_7	v_8	$\ddot{r}_2$	d_8
9	v_5	v_6	v_7	v_8	v_9	$\ddot{r}_2$	d_9
10	v_6	v_7	v_8	v_9	v_{10}	$\ddot{r}_1$	d_{10}
11	v_7	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
$\ddot{r}_1$	$\ddot{v}_{1,4}$	$\ddot{v}_{1,3}$	$\ddot{v}_{1,2}$	$\ddot{v}_{1,1}$	$\ddot{v}_{1,0}$
$\ddot{r}_2$	$\ddot{v}_{2,4}$	$\ddot{v}_{2,3}$	$\ddot{v}_{2,2}$	$\ddot{v}_{2,1}$	$\ddot{v}_{2,0}$

[1] A.A. Cook, G. Misirli, and Z. Fan, 2020, Anomaly Detection for IoT Time-Series Data: A Survey, *IEEE Internet of Things Journal*, v. 7, n. 7, p. 6481–6494.
[2] M. Braei and S. WagnerERRO. Anomaly Detection in Univariate Time-series: A Survey on the State-of-the-Art.
[3] H. Wang, M.J. Bah, and M. Hammad, 2019, Progress in Outlier Detection Techniques: A Survey, *IEEE Access*, v. 7, p. 107964–108000.

Statistical based

- Distribution analysis
  - Analysis of noise – anomaly detection
  - Analysis of window - drift



[1] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, 2019, Learning under Concept Drift: A Review, *IEEE Transactions on Knowledge and Data Engineering*, v. 31, n. 12, p. 2346–2363.
[2] A.S. Iwashita and J.P. Papa, 2019, An Overview on Concept Drift Learning, *IEEE Access*, v. 7, p. 1532–1547.
https://discourse.julialang.org/t/statistic-for-differentiating-two-distributions/31492

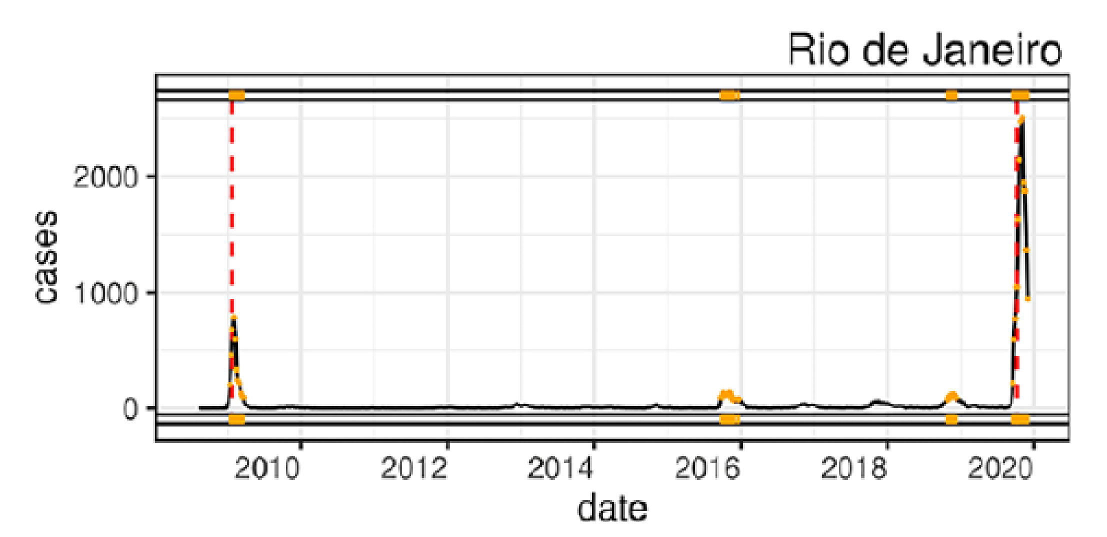
Theory based

- Create a model based on theory
  - Econometric model

$$\bar{y}^{s}_{i,p} = \frac{\sum_{k=1}^{p} t_k}{p} \mid t_k \in seq^{s}_{i,p}(y),\ p \le i \le |y| \tag{3}$$

$$\hat{y}^{s}_{i,p} = \frac{\sum_{k=1}^{p} \alpha_k \cdot t_k}{\sum_{k=1}^{p} \alpha_k} \mid t_k \in seq^{s}_{i,p}(y), \alpha_k = \left(1 - \frac{2}{p+1}\right)^{p-k},\ p \le i \le |y| \tag{4}$$

$$anomaly(y) = \{i\}, \forall i \mid y_i \notin [Q_1(y) - 3 \cdot IQR(y), Q_3(y) + 3 \cdot IQR(y)] \tag{5}$$



New Generation Computing
https://doi.org/10.1007/s00354-021-00125-3

Estimation of COVID-19 Under-Reporting in the Brazilian States Through SARI

Balthazar Paixão[1] · Laís Baroni[1] · Marcel Pedroso[2] · Rebecca Salles[1] · Luciana Escobar[1] · Carlos de Sousa[2] · Raphael de Freitas Saldanha[2] · Jorge Soares[1] · Rafaelli Coutinho[1] · Fabio Porto[3] · Eduardo Ogasawara[1]

Received: 19 December 2020 / Accepted: 4 March 2021

Abstract
Due to its impact, COVID-19 has been stressing the academy to search for curing, mitigating, or controlling it. It is believed that under-reporting is a relevant factor in determining the actual mortality rate and, if not considered, can cause significant misinformation. Therefore, this work aims to estimate the under-reporting of cases and deaths of COVID-19 in Brazilian states using data from the InfoGripe. InfoGripe targets notifications of Severe Acute Respiratory Infection (SARI). The methodology is based on the combination of data analytics (event detection methods) and time series modeling (inertia and novelty concepts) over hospitalized SARI cases. The estimate of real cases of the disease, called novelty, is calculated by comparing the difference in SARI cases in 2020 (after COVID-19) with the total expected cases in recent years (2016–2019). The expected cases are derived from a seasonal exponential moving average. The results show that under-reporting rates vary significantly between states and that there are no general patterns for states in the same region in Brazil. The states of Minas Gerais and Mato Grosso have the highest rates of under-reporting of cases. The rate of under-reporting of deaths is high in the Rio Grande do Sul and the Minas Gerais. This work can be highlighted for the combination of data analytics and time series modeling. Our calculation of under-reporting rates based on SARI is conservative and better characterized by deaths than for cases.

Keywords COVID-19 · Under-reporting · SARI · Time series modeling · Data analytics

Laís Baroni

1 Federal Center for Technological Education of Rio de Janeiro, CEFET/RJ, Rio de Janeiro, Brazil

2 Oswaldo Cruz Foundation, Fiocruz, Rio de Janeiro, Brazil

3 National Laboratory of Scientific Computing, LNCC, Rio de Janeiro, Brazil

Published online: 14 March 2021

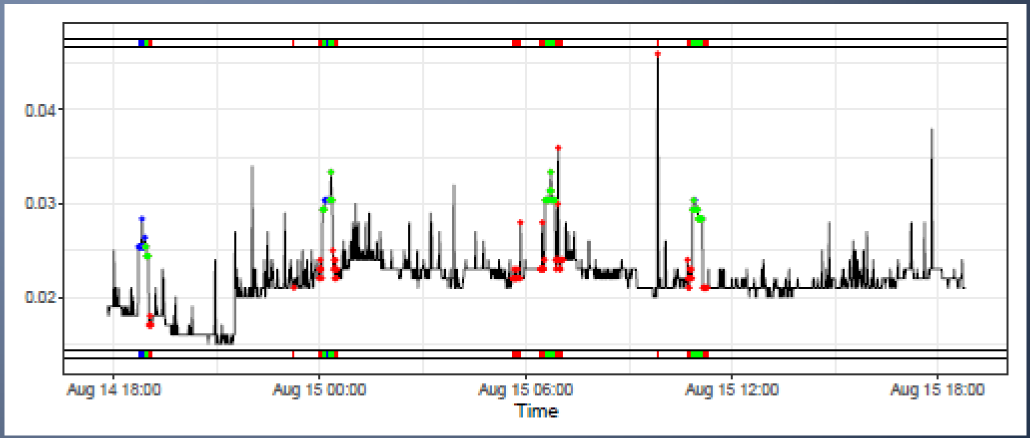
[1] B. Paixão, L. Baroni, M. Pedroso, R. Salles, L. Escobar, C. de Sousa, R. de Freitas Saldanha, J. Soares, R. Coutinho, et al., 2021, Estimation of COVID-19 Under-Reporting in the Brazilian States Through SARI, *New Generation Computing*, v. 39, n. 3–4, p. 623–645.

Accurateness

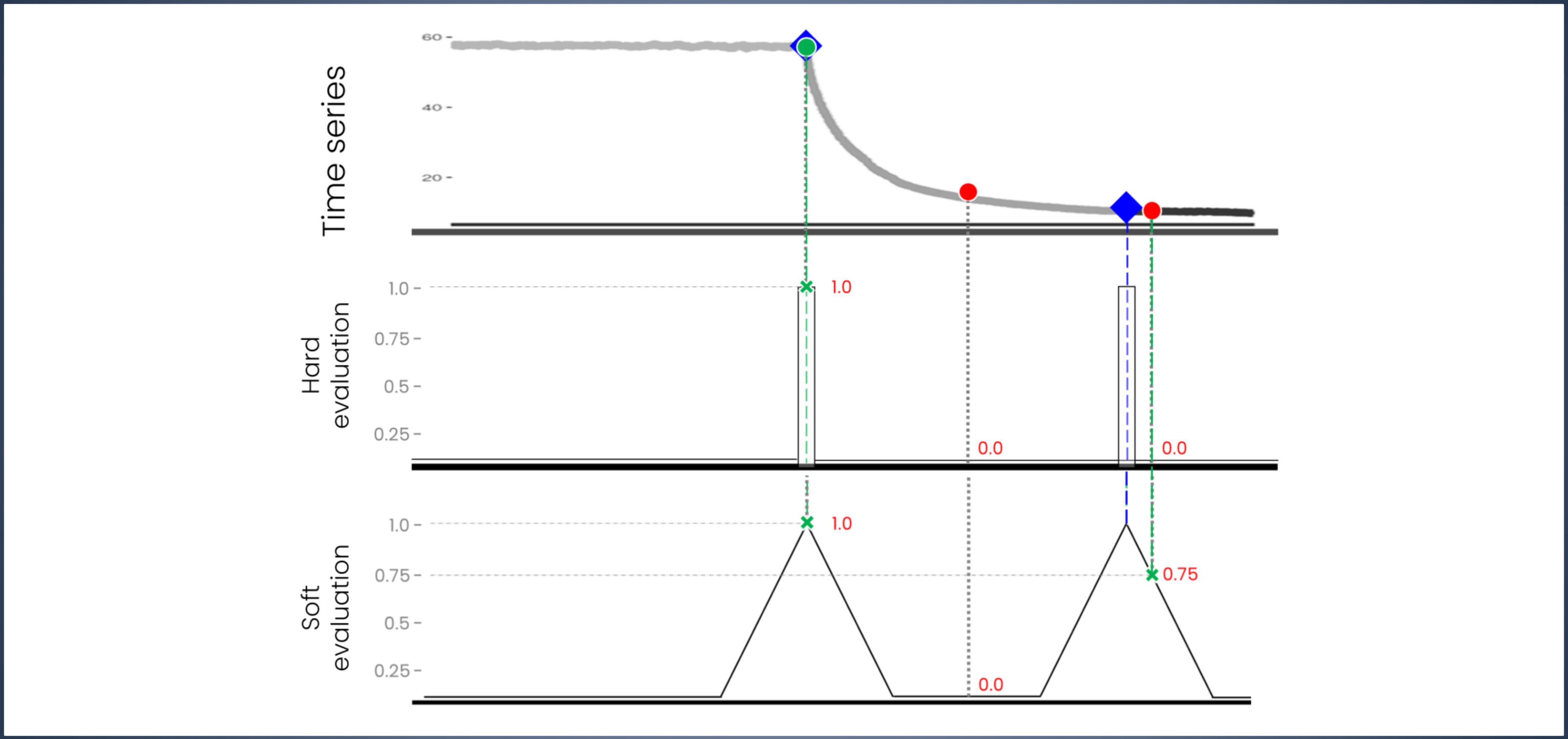
- Classifier Accuracy: percentage of test set tuples that are correctly classified
  - $accurary = \frac{TP+TN}{All}$
  - $precision = \frac{TP}{TP+FP}$
  - $recall = \frac{TP}{TP+FN}$
  - $F_1 = \frac{2 \cdot precision \cdot recall}{precision+recall}$
  - *ROC Curve*

Confusion Matrix (CM)

Predicted Actual	Ê	¬Ê
E	TP	FN
¬E	FP	TN



Time tolerance in detection



[1] R. Salles, J. Lima, R. Coutinho, E. Pacitti, F. Masseglia, R. Akbarinia, C. Chen, J. Garibaldi, F. Porto, et al.ERRO. SoftED: Metrics for Soft Evaluation of Time Series Event Detection.

Expensiveness

- Elapsed time
- Time constraints for online detection
  - Drift
  - Incremental learning

Data Analytics Lab Team

Doutorado

Lais Baroni (CEFET/RJ)

Leonardo Carvalho (CEFET/RJ)

Rebecca Salles (CEFET/RJ)

Mestrado

Antônio Mello (CEFET/RJ)

Arthur Garcia (CEFET/RJ)

Cristiane Gea (CEFET/RJ)

Diego Salles (CEFET/RJ)

Janio Lima (CEFET/RJ)

Jéssica de Souza (CEFET/RJ)

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

Laboratório
Nacional de
Computação
Científica

TIME SERIES
EVENT DETECTION

CEFET/RJ

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