CEFET/RJ

EVENT DETECTION

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

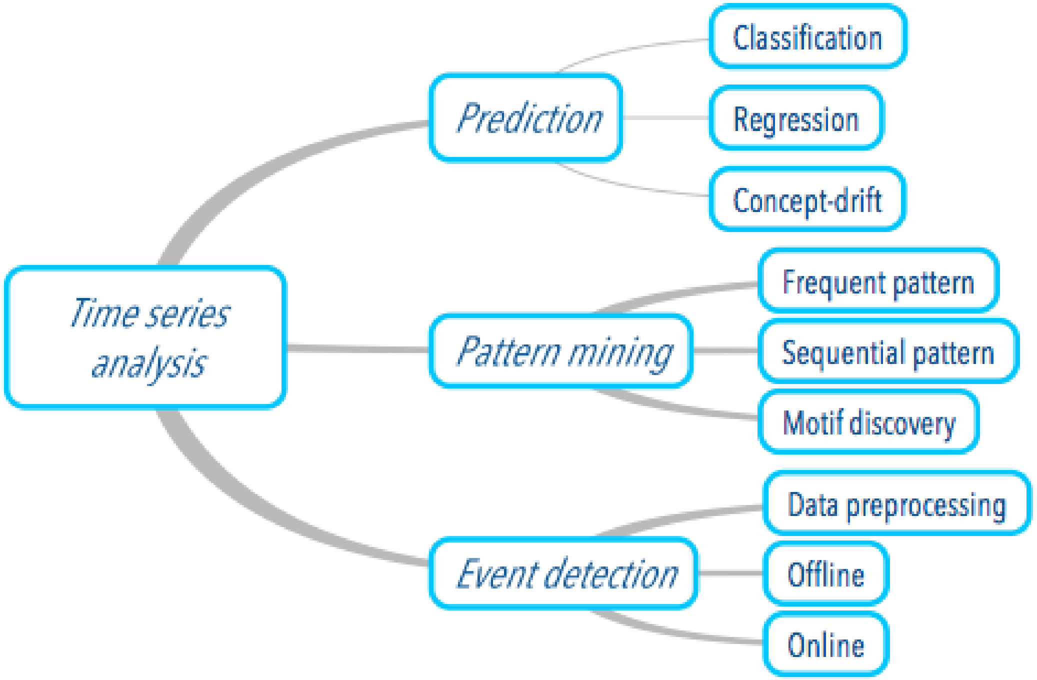
Biografia

- Doutorado em Engenharia de Sistemas e Computação (COPPE/UFRJ) em 2011
- Docente da EIC - CEFET/RJ
  - Departamento de Ciência da Computação
  - Coordenação do Curso Técnico de Informática
- Docente permanente
  - Programa de Pós-graduação em Ciência da Computação (PPCIC)
  - Programa de Pós-graduação em Eng. de Produção e Sistemas (PPPRO)
- Membro da IEEE, SBC, ACM e INNS

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

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

Research Themes



Background

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

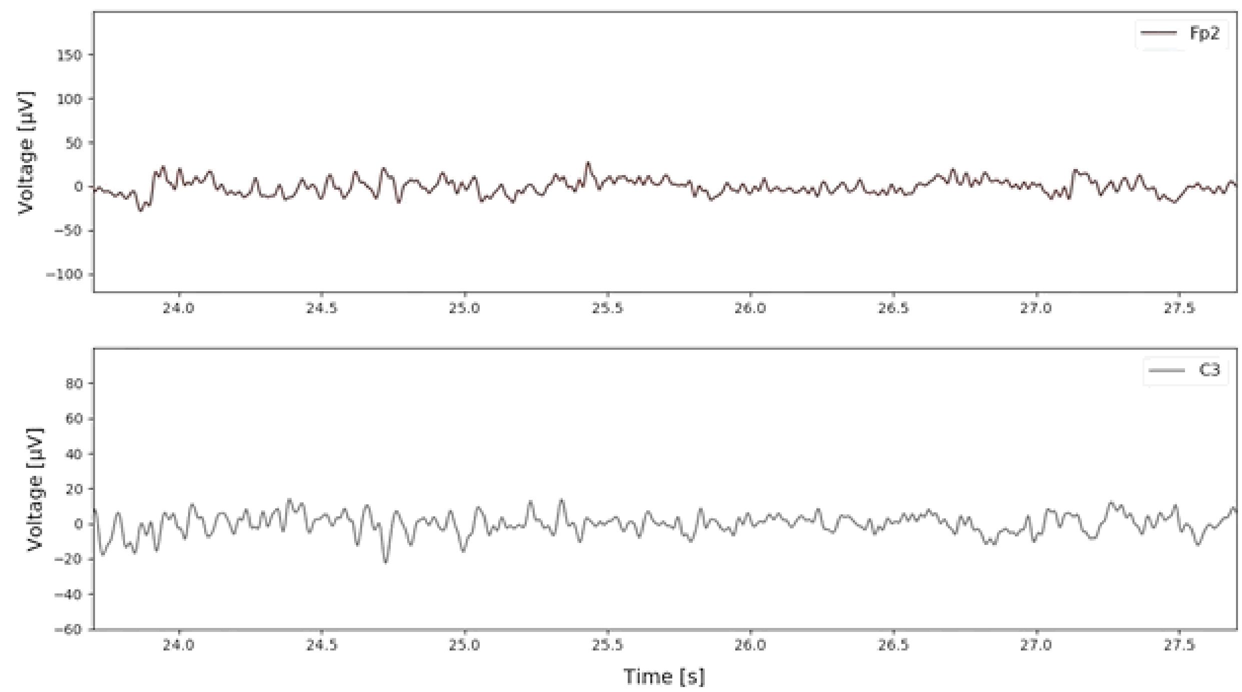
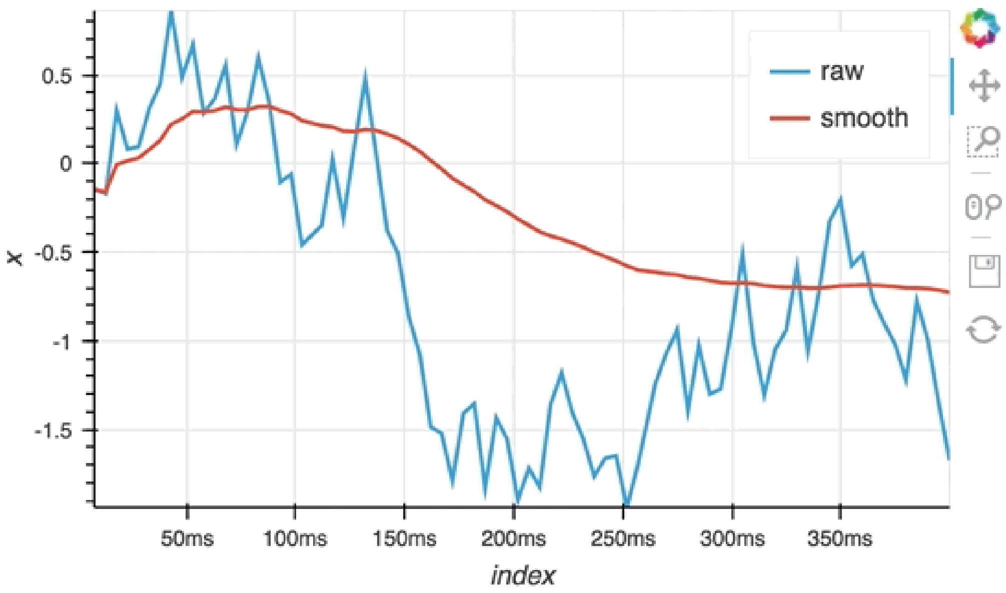


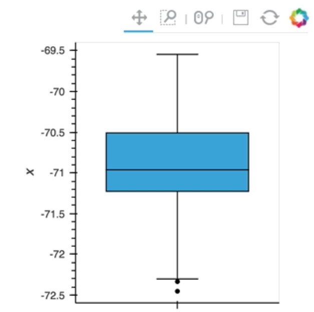
Image source: https://towardsdatascience.com/dynamic-replay-of-time-series-data-819e27212b4b

Time series – online analysis

- Statistical properties may vary over time in streaming data



#	index	x
0	2017-10-27 19:43:04	-70.87262710547351
1	2017-10-27 19:43:04	-71.00788295730518
2	2017-10-27 19:43:04	-71.01297392873352
3	2017-10-27 19:43:04	-71.4148637783796
4	2017-10-27 19:43:04	-71.60890969520968
5	2017-10-27 19:43:04	-71.32610485802545
6	2017-10-27 19:43:04	-71.1935680343768
7	2017-10-27 19:43:04	-70.76579342173753
8	2017-10-27 19:43:04	-70.59743524950701
9	2017-10-27 19:43:04	-70.99300214112863



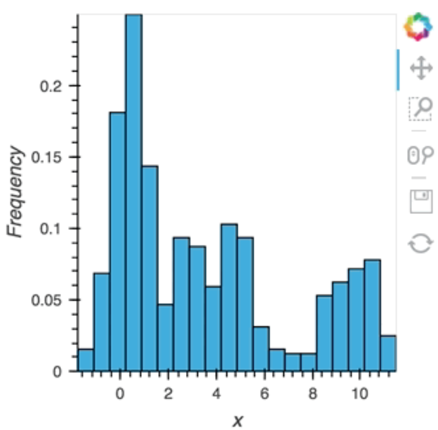


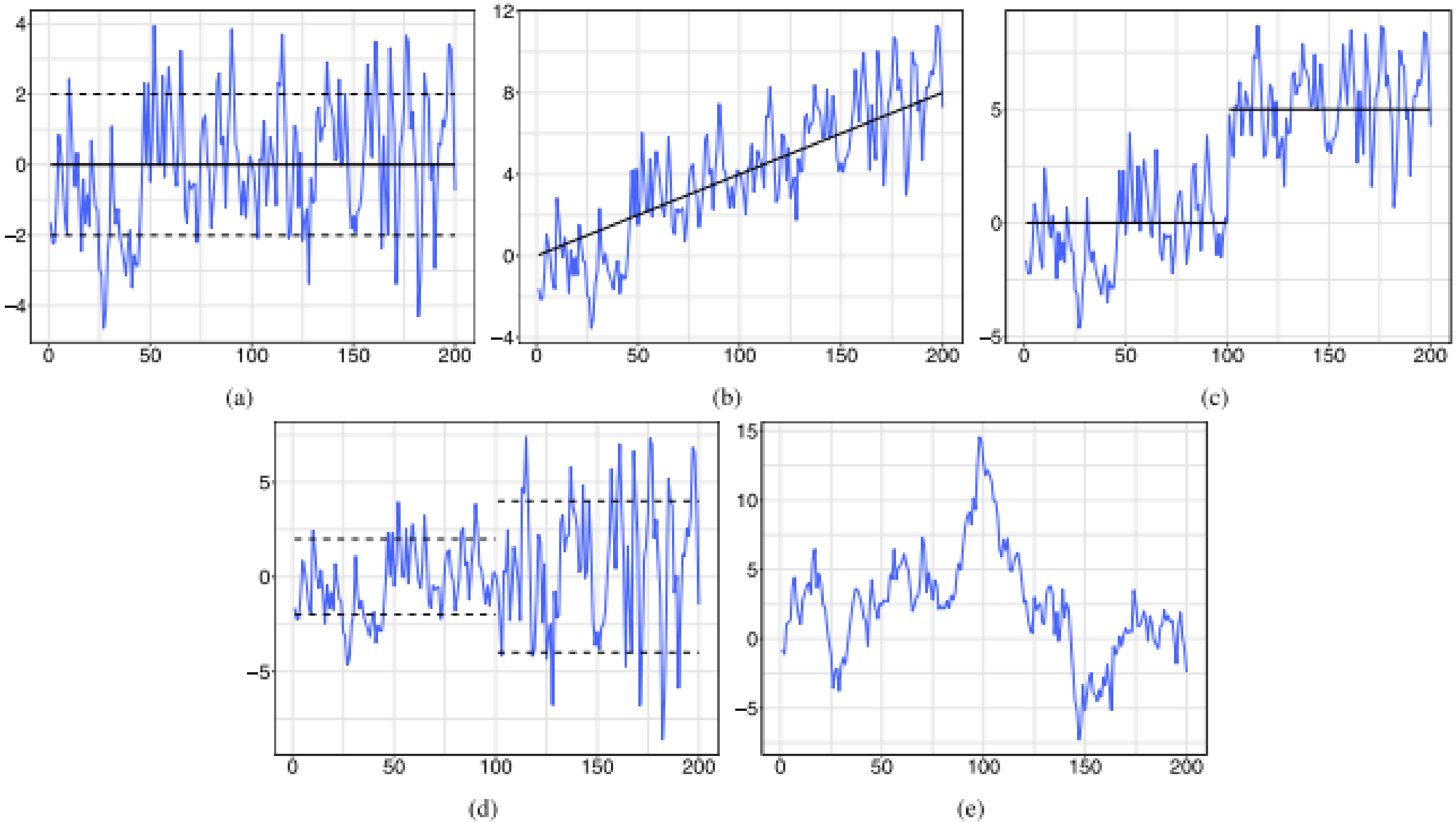
Image source: http://holoviews.org/user_guide/Streaming_Data.html

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

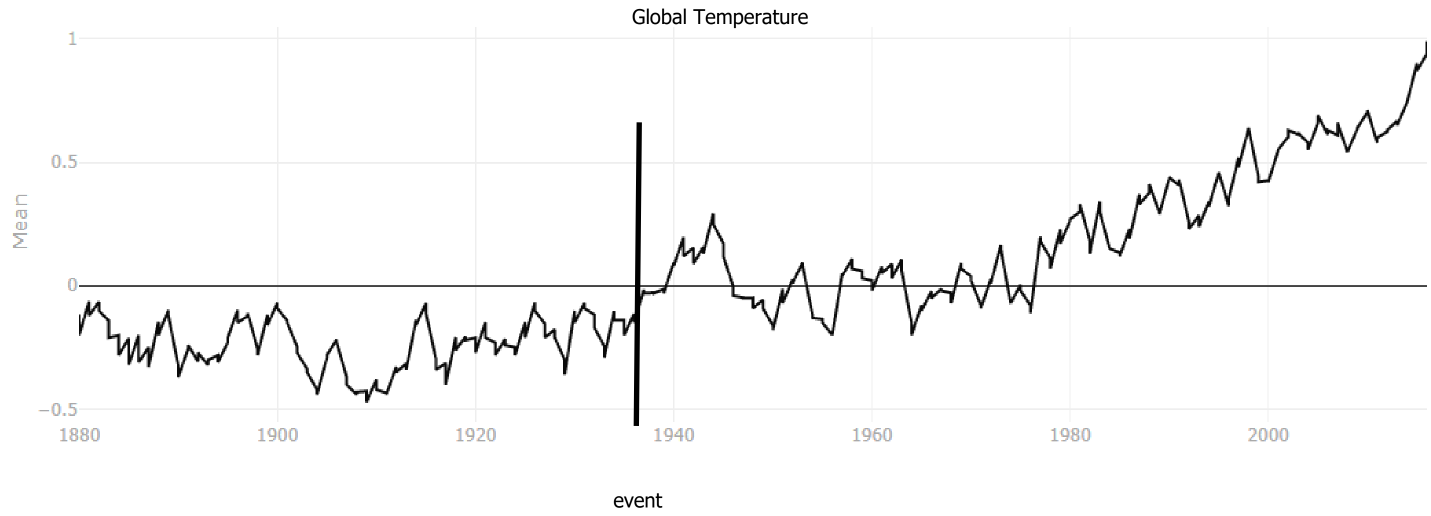


R. Salles, K. Belloze, F. Porto, P. H. Gonzalez, e E. Ogasawara, “Nonstationary time series transformation methods: An experimental review”, Knowledge-Based Systems, nov. 2018.

Event detection

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)



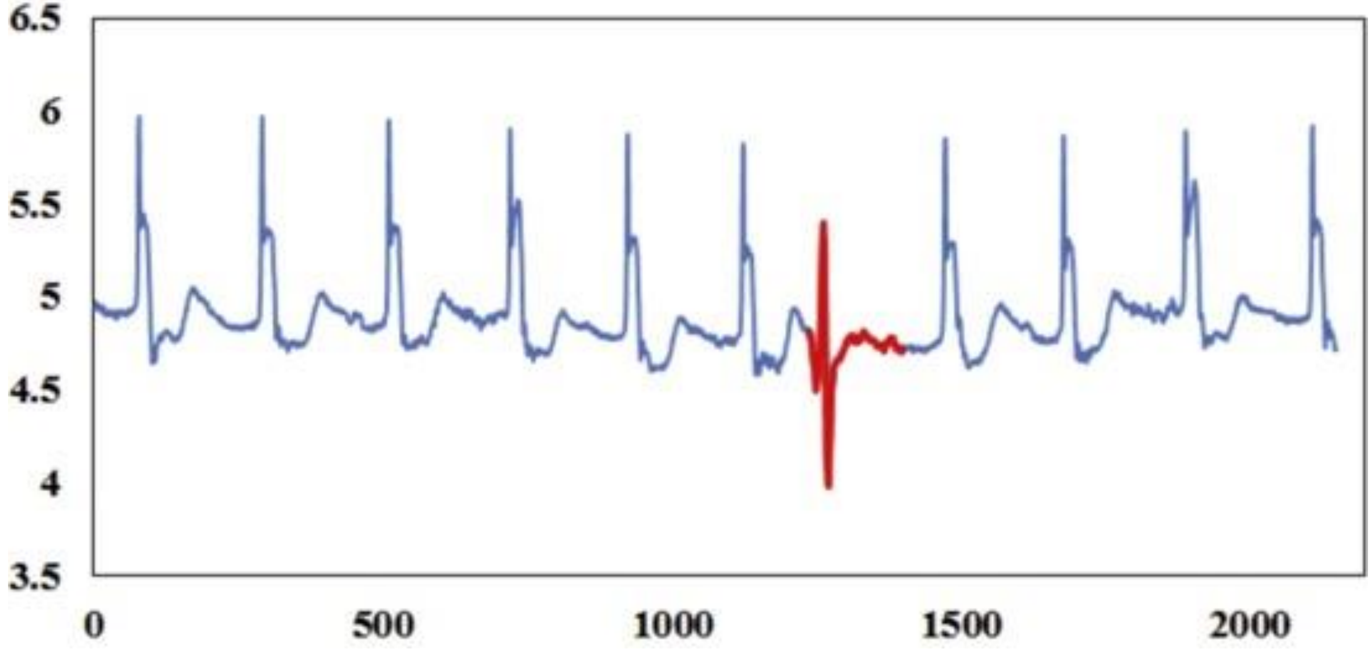
[1] V. Guralnik and J. Srivastava, 1999, Event Detection from Time Series Data, In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, p. 33–42

Image source: World Global Temperature, https://datahub.io/core/global-temp

Anomalies

- A pattern or observation that do not conform to expected behavior [1]
- It can be categorized as punctual, contextual or collective

(*)

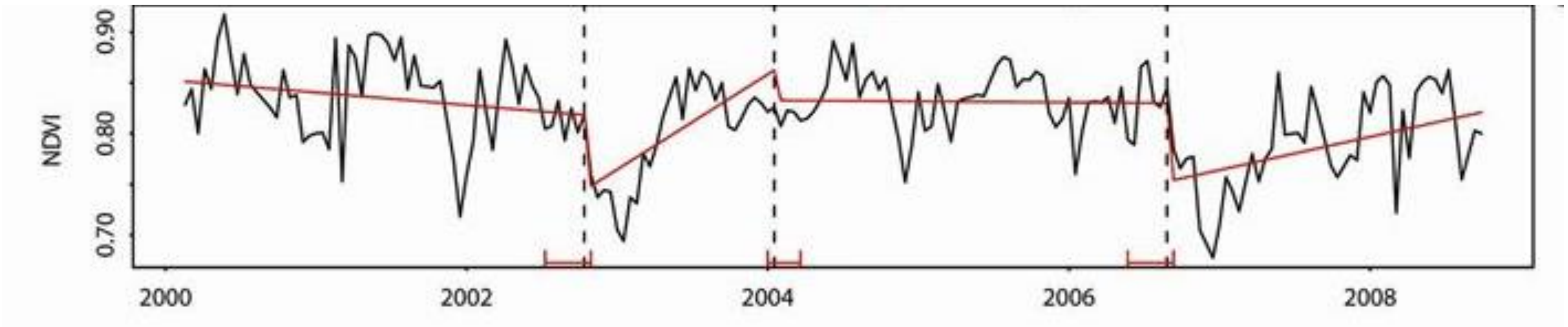


[1]] V. Chandola, A. Banerjee, e V. Kumar, 2009, Anomaly detection: A survey, ACM Computing Surveys, v. 41, n. 3

(*) In this example, it can also be classified as a discord

Change Points

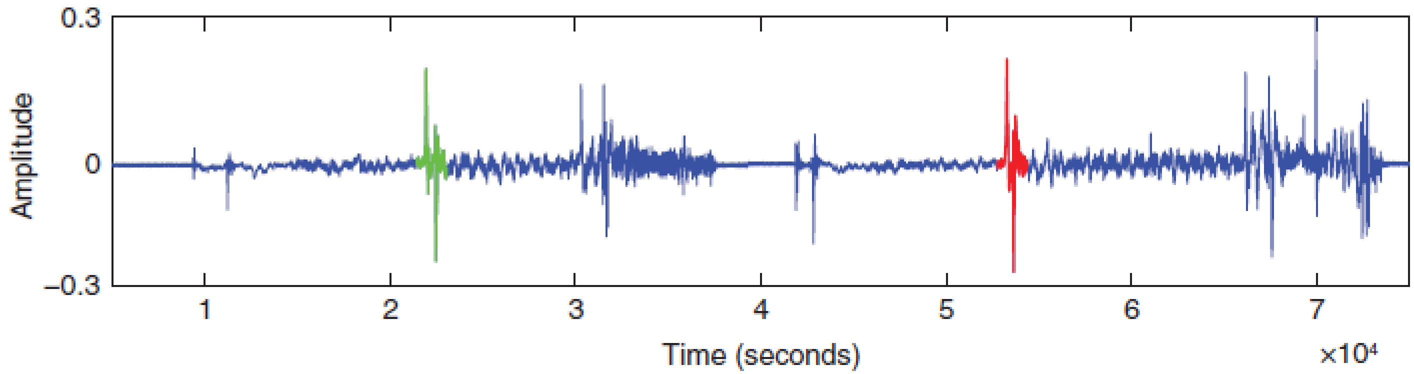
- Points (or time intervals) that mark significant change in time series behavior [1]
- They separate different states in the process that generates the time series



[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 [1,2,3]



How to do it in non-stationarity 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.

Event detection

- An event can represent a phenomenon with a specific meaning defined in a certain domain
- Event detection is the process of finding events
- It is a basic function in surveillance and monitoring systems
- Example of applications:

Eletronic IT security

Healthcare

Medical diagnostics and monitoring

Industrial monitoring & Damage detection

Sensor Networks

[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.
[3] H. Wang, M.J. Bah, e M. Hammad, 2019, Progress in Outlier Detection Techniques: A Survey, IEEE Access, v. 7, p. 107964–108000.

Importance of event detection

[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.

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.

Anomaly detection (distribution analysis)

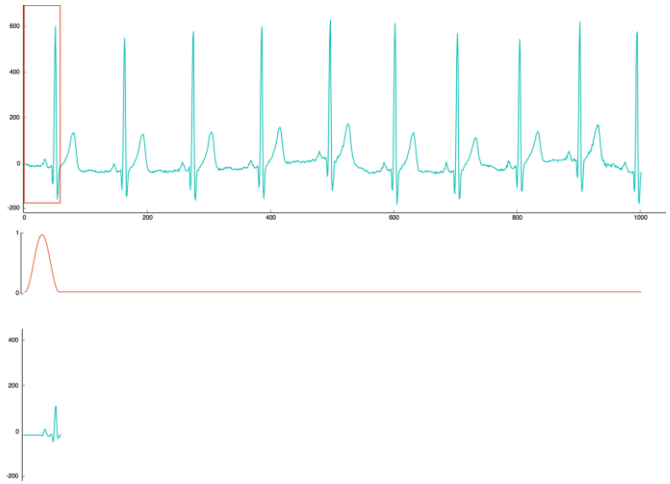
- Statistical analysis
  - Differentiation (backshift operator)
  - Residuals from moving average
  - Residuals from filters (Kalman)
- Model adjustment
  - Residuals from decomposed signal
  - Residuals from linear models (regression)
  - Residuals from autoregressive models (ARIMA)
  - Residuals from volatility models (GARCH)
  - Residuals from machine learning models
- Clustering of subsequences
  - Distribution analysis over difference between subsequences and centroids
  - DBScan
- Time series decomposition
  - Trend
  - Seasonal
  - Fourier transform
  - Wavelets
  - IMF - intrinsic mode function
  - Hilbert-Huang transform

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

Change point detection

- Seminal change point [1]
- Change Finder [2]

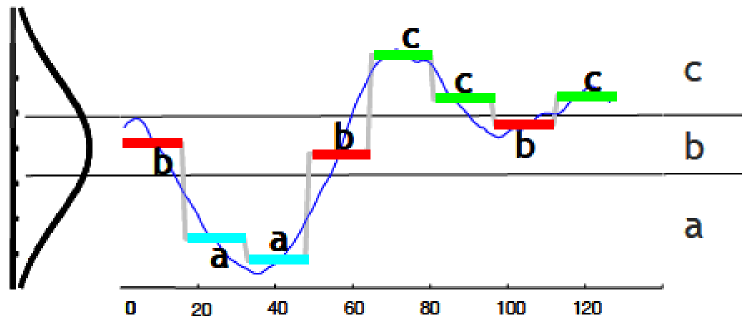
Windowed approach

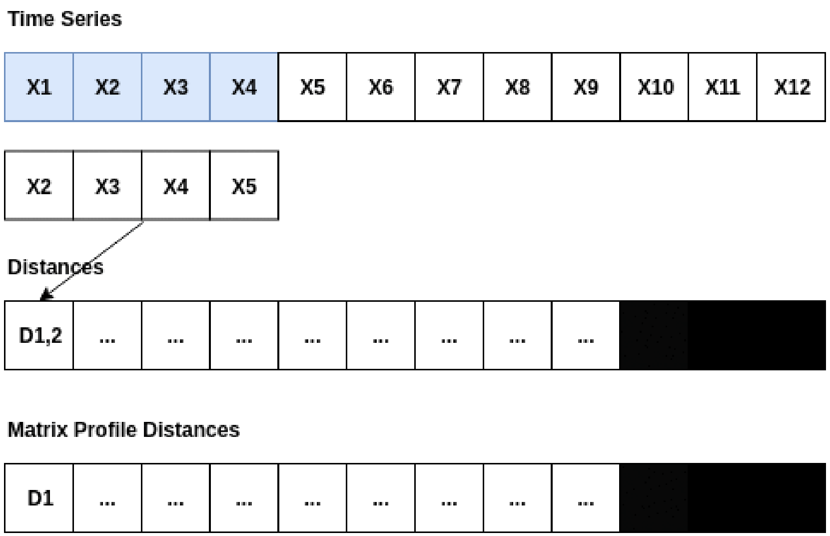


[1] J.-I. Takeuchi and K. Yamanishi, 2006, A unifying framework for detecting outliers and change points from time series, IEEE Transactions on Knowledge and Data Engineering, v. 18, n. 4, p. 482–492.
[2] V. Guralnik and J. Srivastava, 1999, Event Detection from Time Series Data, In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, p. 33–42
Image source: https://towardsdatascience.com/real-time-time-series-anomaly-detection-981cf1e1ca13

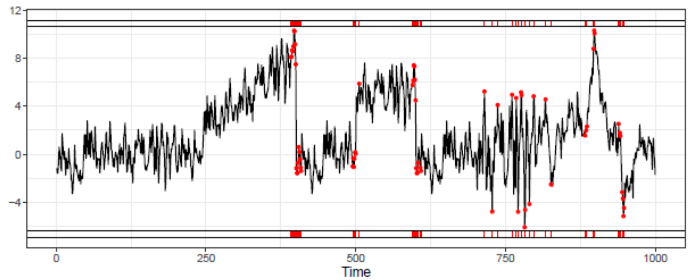
Motif discovery

- Indexing
  - Discretization
  - SAX [1]
- Brute force
- Hash-based (random projection) [2]
- Matrix profile [3]

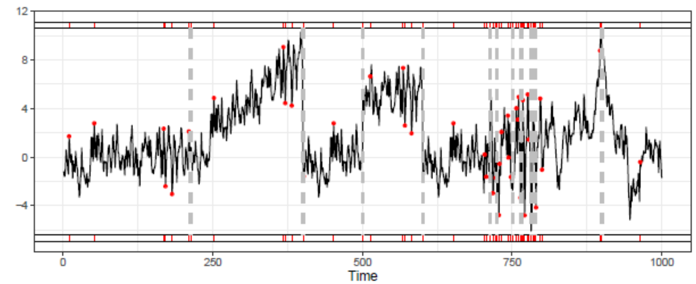




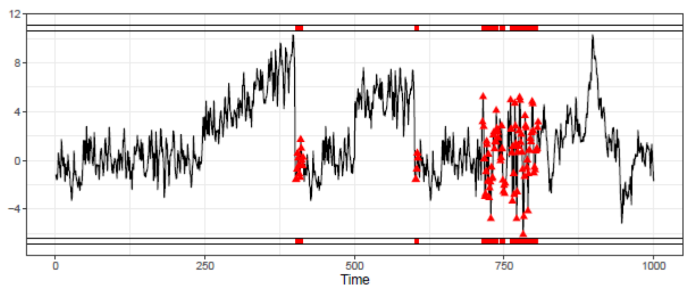
[1] J. Lin, E. Keogh, L. Wei, and S. Lonardi, "Experiencing SAX: a novel symbolic representation of time series," Data Mining and Knowledge Discovery, vol. 15, no. 2, pp. 107–144, 2007
[2] A. Mueen, "Time series motif discovery: Dimensions and applications," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 4, no. 2, pp. 152–159, 2014
[3] M. Linardi, Y. Zhu, T. Palpanas, and E. Keogh, 2020, Matrix profile goes MAD: variable-length motif and discord discovery in data series, *Data Mining and Knowledge Discovery*, v. 34, n. 4, p. 1022–1071.
(*) https://towardsdatascience.com/introduction-to-matrix-profiles-5568f3375d90



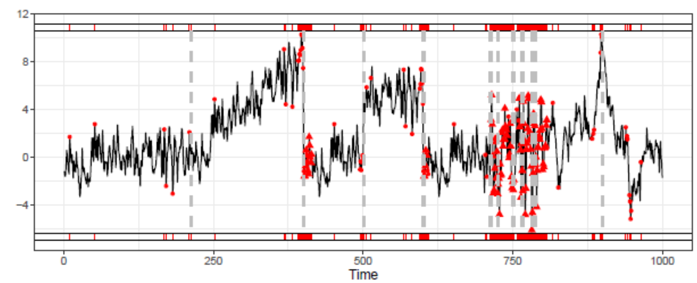
Method A: trend anomalies



Method B: trend anomalies & change points



Method C: volatility anomalies



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

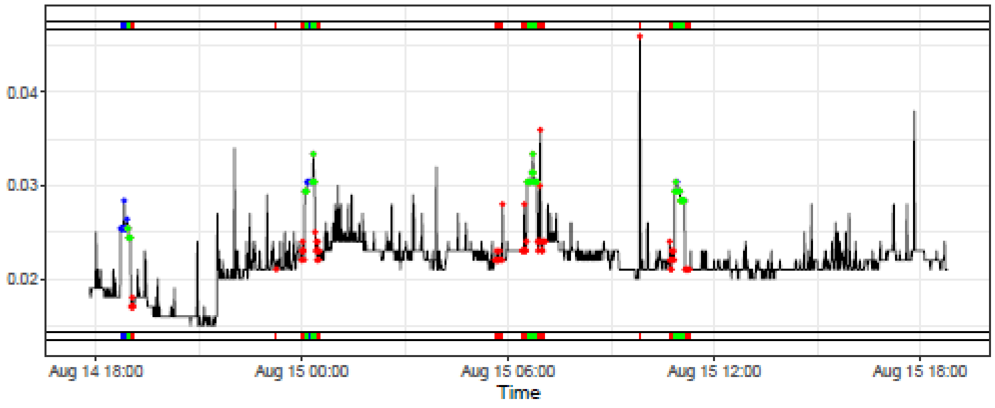
[1] R. Salles, L. Escobar, L. Baroni, R. Zorrilla, A. Ziviani, V. Kreischer, F. Delicato, P. Pires, L. Maia, et al., 2020, Um framework para integração e análise de métodos de detecção de eventos em séries temporais, In: *Anais do Simpósio Brasileiro de Banco de Dados (SBBD)*

Metrics for event detection

- 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*

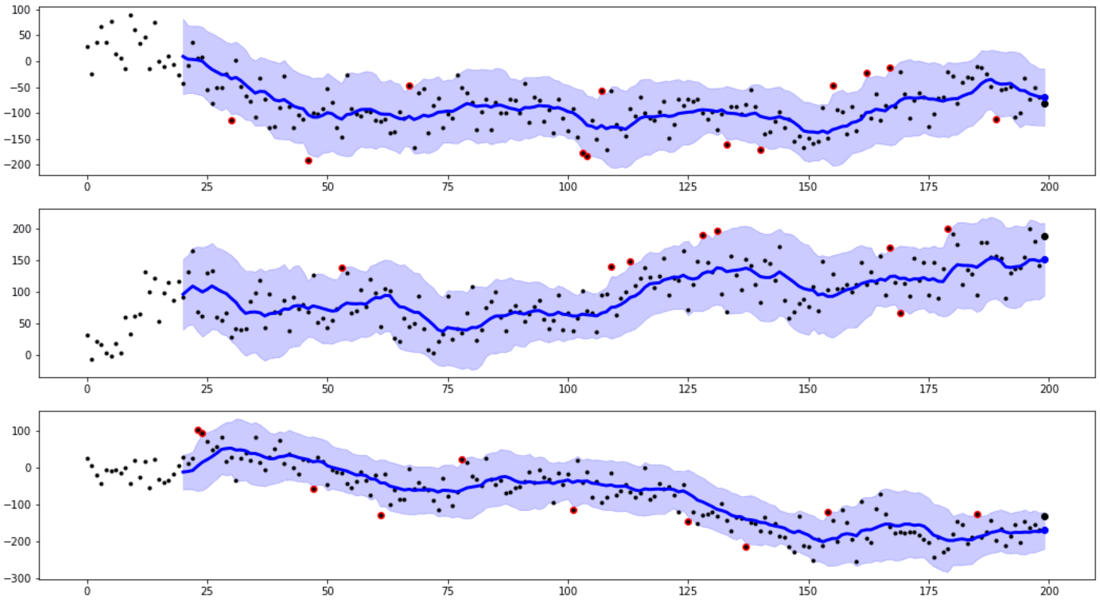
Confusion Matrix (CM)

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

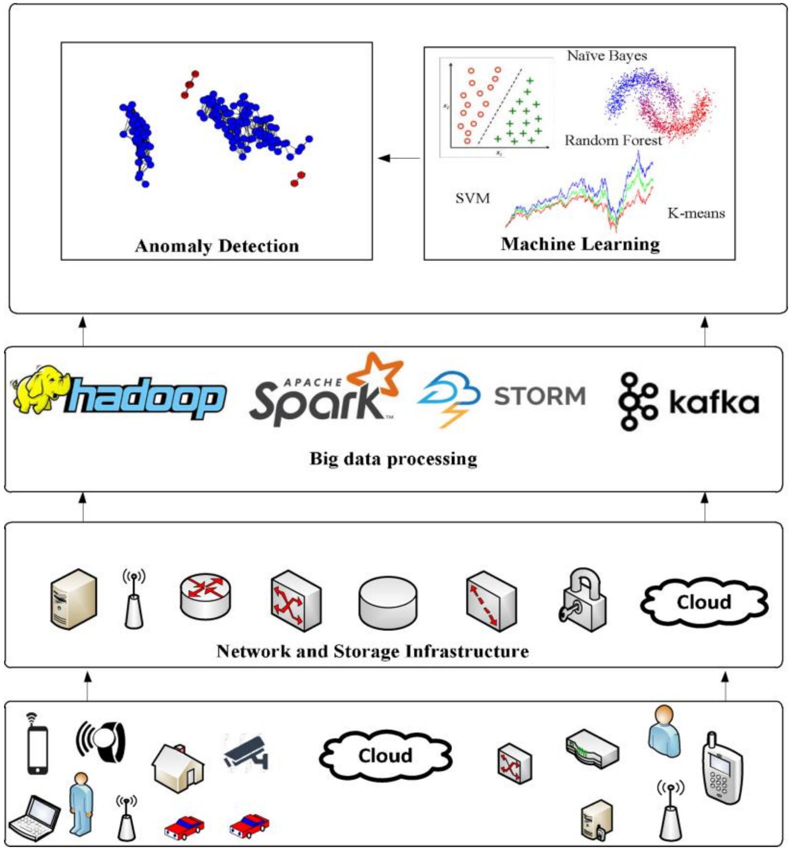


Online event detection

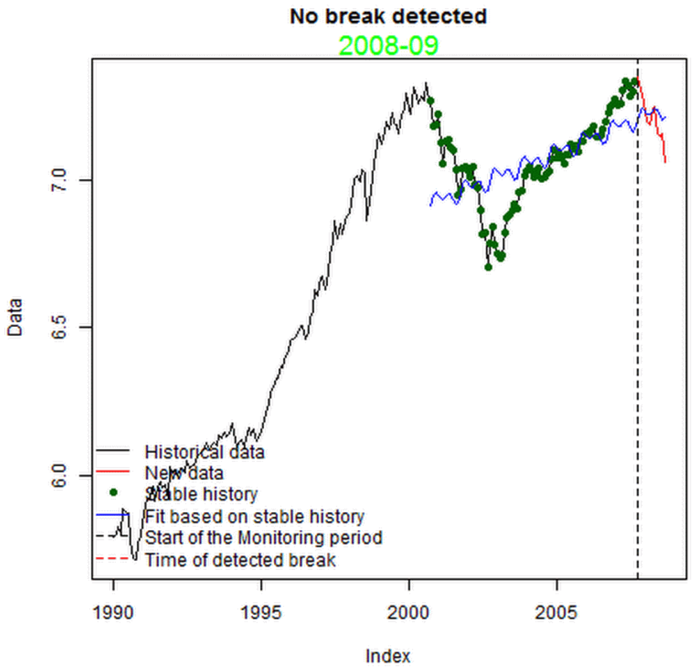
- Handles streaming time series

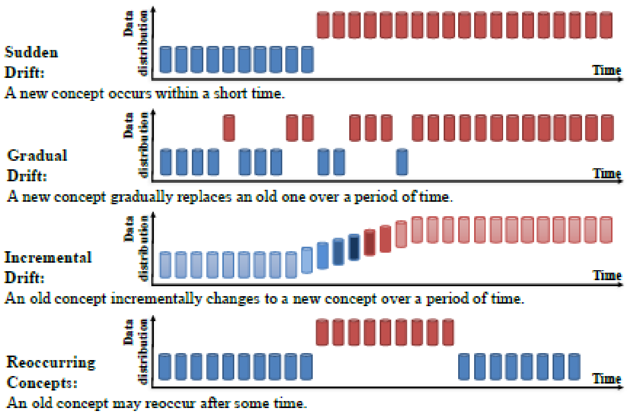


Online event detection infrastructure



- Detection occurs incrementally





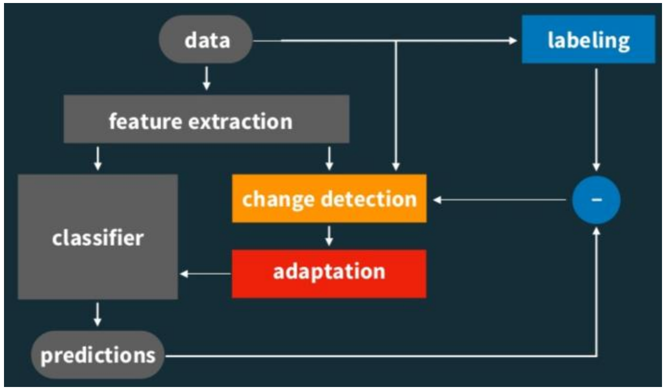


Image source: https://towardsdatascience.com/real-time-time-series-anomaly-detection-981cf1e1ca13
[1] G. Ditzler, M. Roveri, C. Alippi, e R. Polikar, 2015, Learning in Nonstationary Environments: A Survey, *IEEE Computational Intelligence Magazine*, v. 10, n. 4, p. 12–25.
[2] https://youtu.be/woRmeGvOaz4

Online event detection challenges
(when to adapt)

Adaptability:

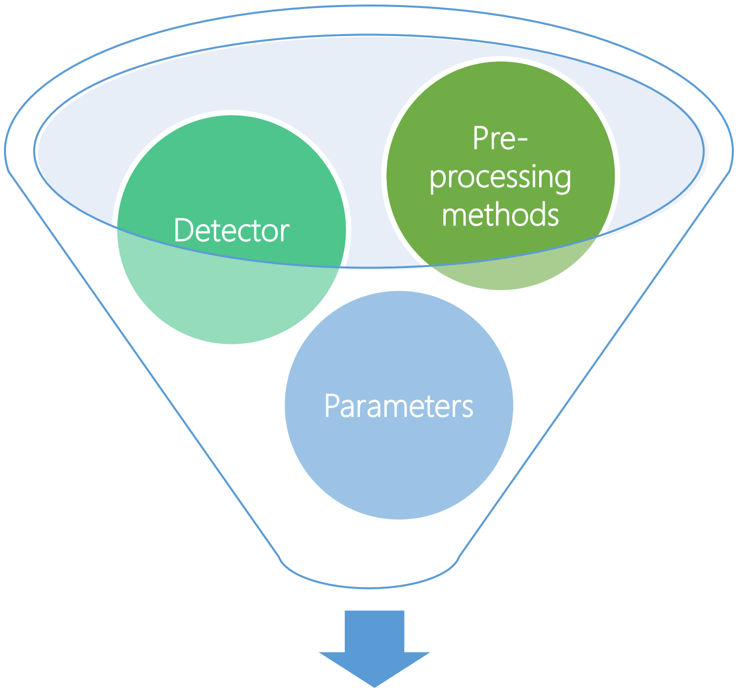
Incremental learning

Plasticity and Stability Dilemma

[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.

Online event detection challenges (too many methods)

Myriad of event detection methods (detectors)



Detection performance

Choice of appropriate detectors/parameters for event detection is a challenge

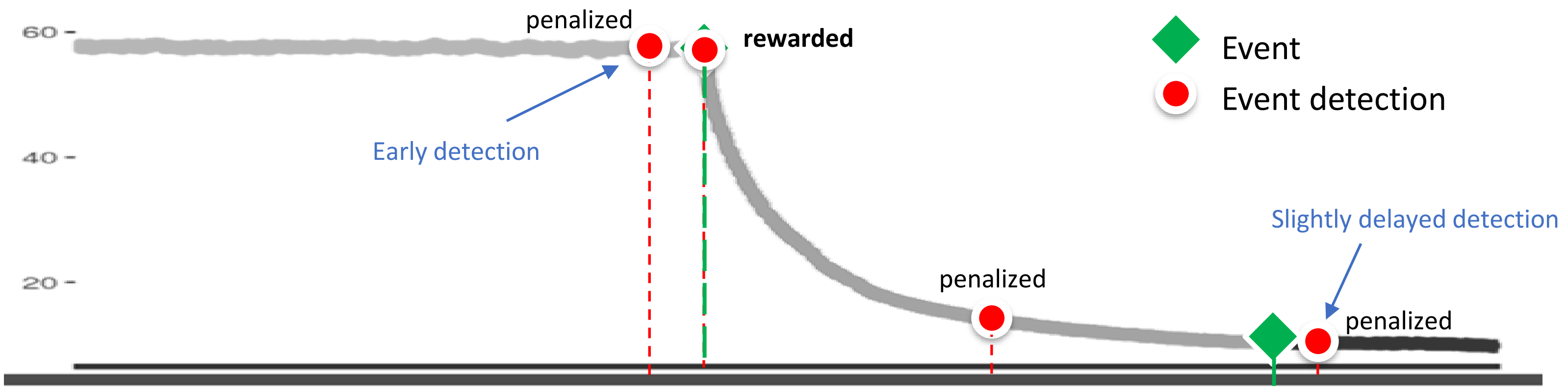
Directly related to initial assumptions about the behavior and statistical properties of data

The nature of the events observed is often unknown

Detectors specialized in a type of event may disregard the occurrence of another, or even misidentify them

Online event detection challenges (metrics)

- Traditional scoring methods, such as precision and recall, don't suffice for evaluating online event detection performance.
  - They do not incorporate time and do not reward early detection.
  - True positives are rewarded. All other results are "harshly" and equally punished.



[1] Lavin, A., & Ahmad, S. (2015, December). Evaluating Real-Time Anomaly Detection Algorithms--The Numenta Anomaly Benchmark. In 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA) (pp. 38-44). IEEE.

[2] Singh, N., & Olinsky, C. (2017, May). Demystifying Numenta anomaly benchmark. In 2017 International Joint Conference on Neural Networks (IJCNN) (pp. 1570-1577). IEEE.

Ongoing Studies

Ongoing research

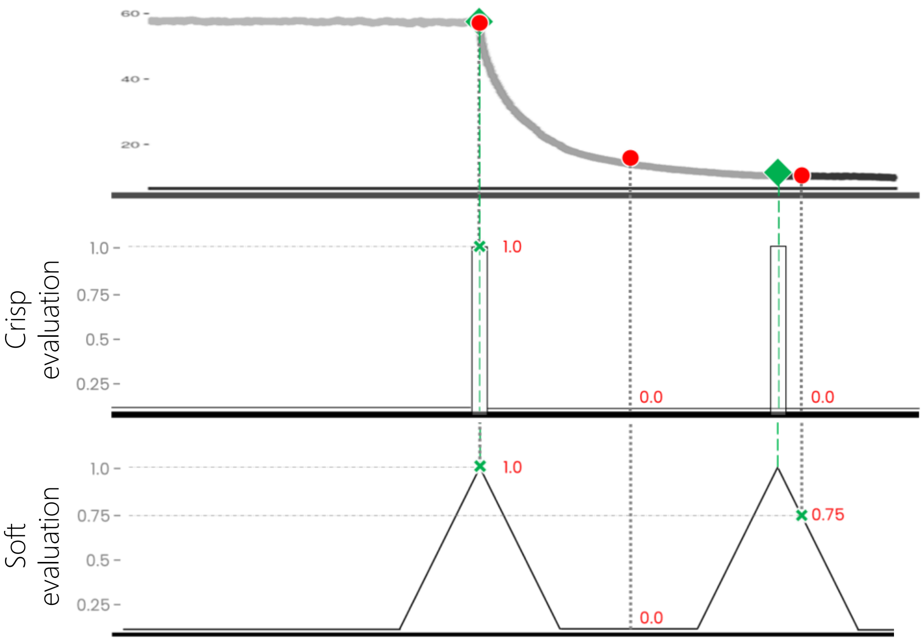
- Novel soft metric for measuring event detection
- Exploring different methods and ensemble approaches
  - Harbinger: Identified from systematic review
  - RED: Exploring nonstationary transformation methods and decomposition
- Exploring tools for real-time big data processing
  - Harbinger Nimbus: Microsoft Azure, Apache Spark (sparklyr), and Apache Kafka
- Applications
  - Event detection for covid-19 underreport
  - Event detection for neonatal mortality (Fiocruz)
    - Politics: hospitals with breastfeeding supports [2]
  - Event detection for oil well drilling and exploration (Petrobras)
    - Stuck pipe [1] and False kicks (variations of fluid used during drilling)
  - Event detection in finance data

[1] MITCHELL, F. R.; LAKE, W. L. Petroleum engineering handbook volume II. Richardson: SPE, 2006.
[2] R. Pérez-Escamilla, J.L. Martinez, e S. Segura-Pérez, 2016, Impact of the Baby-friendly Hospital Initiative on breastfeeding and child health outcomes: a systematic review, Maternal and Child Nutrition, v. 12, n. 3, p. 402–417.

Novel soft metric for measuring event detection

- New scoring methods and classification metrics based on fuzzy logic
- Define membership of an event detection to an event related soft class
- Soft version of all know metrics
  - accuracy, precision, recall, f1

Time series



Exploring different methods and ensemble approaches

- Allow integration and analysis of event detection methods [1]
- Allow the identification of different types of events
- Allow the combination of different personalized detection methods
- Enable a comparative analysis of their detections

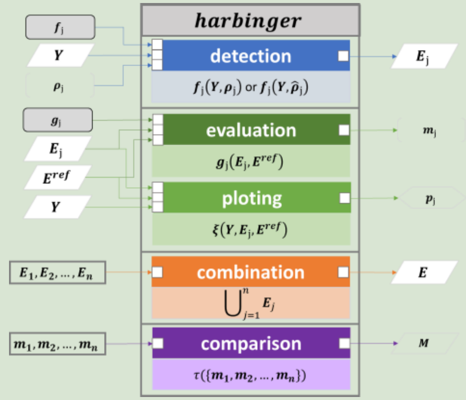
Harbinger

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

Published in

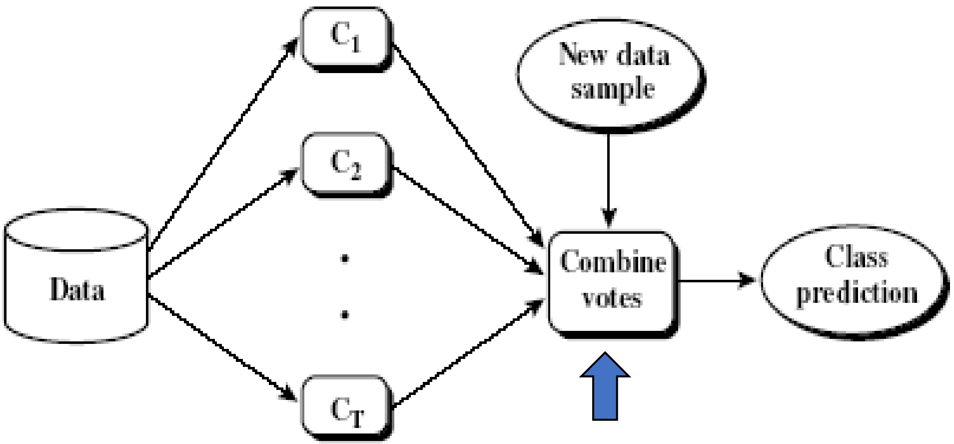
SBBD 2020

Developed in



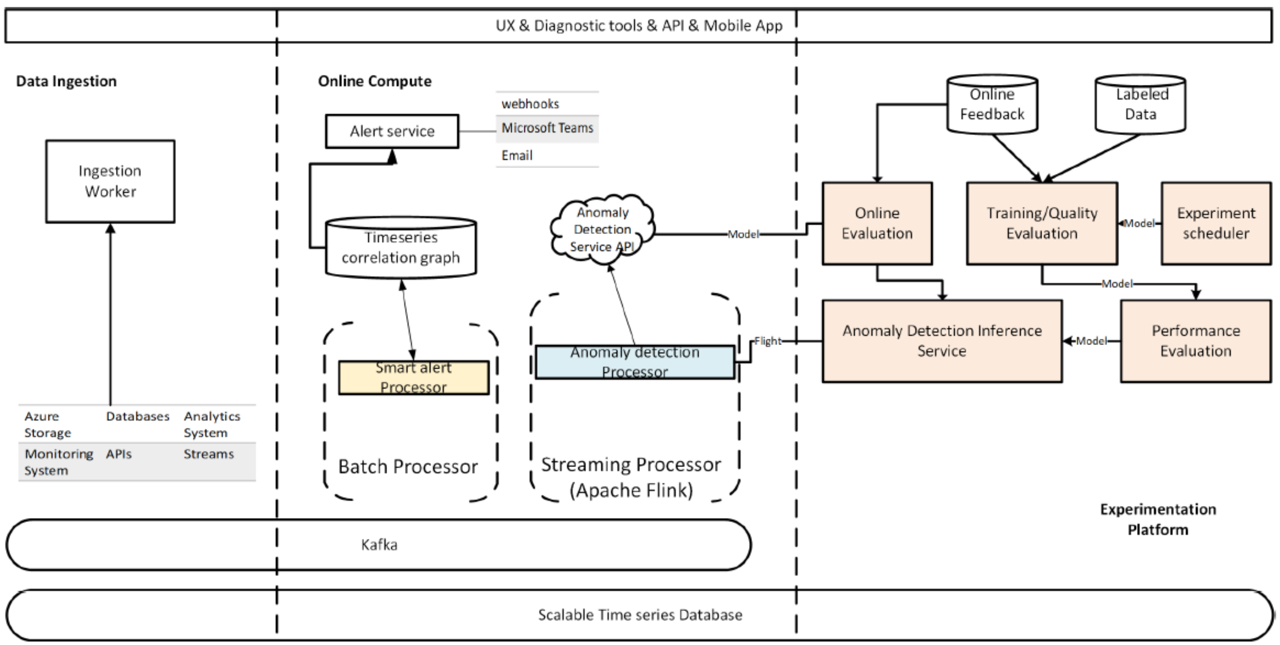
Exploring different methods and ensemble approaches

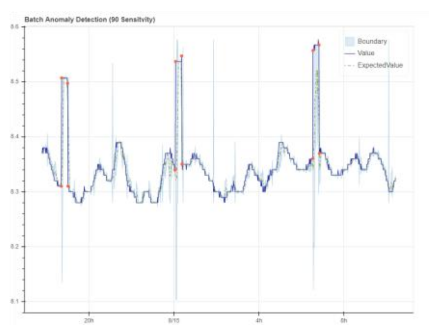
- Ensemble models
  - Makes no assumptions and combines the results of many detection methods
  - Takes advantage of each algorithm
  - Allows parallel computing to improve efficiency
  - Generally, perform better than isolated methods



Harbinger Nimbus

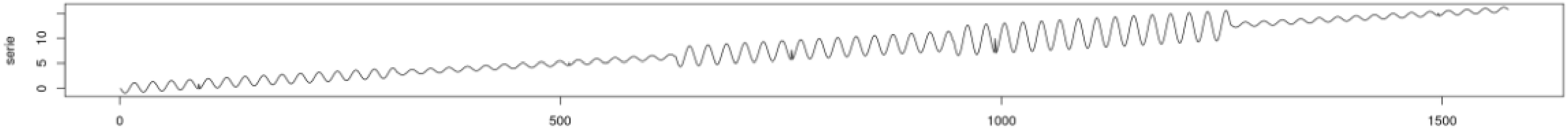
- Integration of Harbinger with Microsoft Azure Event Detection System
- Online event detection



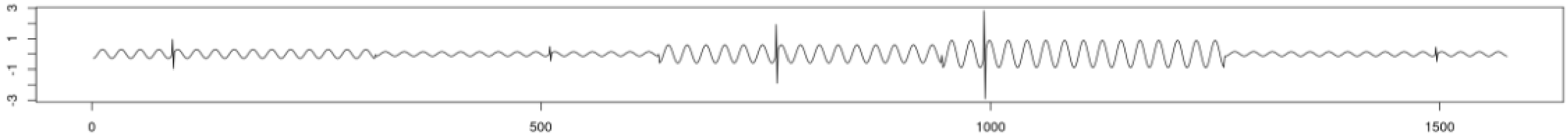


H. Ren, B. Xu, Y. Wang, C. Yi, C. Huang, X. Kou, T. Xing, M. Yang, J. Tong, and Q. Zhang. Time-series anomaly detection service at microsoft. InProceedings of the ACM SIGKDD In-ternational Conference on Knowledge Discovery and Data Mining, pages 3009–3017, 2019.

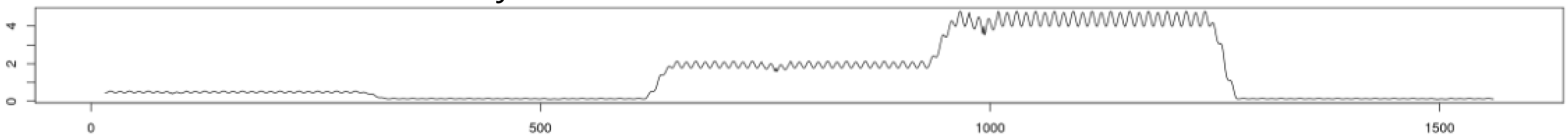
Serie (y_t)



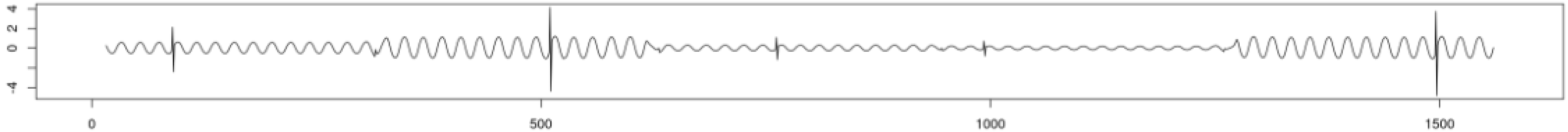
(∇_t) - Differentiation



(σ^2) – Instant volatility



(φ_t) – Harmonic time series



- SARI
- Econometric model
  - Based on inertial concepts and novelty
  - Parameters computed by event detection (change points)

$$y_i - \hat{y}^s_{i-s,p} - \epsilon_i = 0 \quad (7)$$

$$y_i - \hat{y}^s_{i-s,p} - \eta_i - \hat{\epsilon}_i = 0,\ \hat{\epsilon}_i \approx \overline{\epsilon},\ \hat{\epsilon}_i \in [\overline{\epsilon}_{min} - \overline{\epsilon}_{max}] \quad (8)$$


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] · Lais 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.

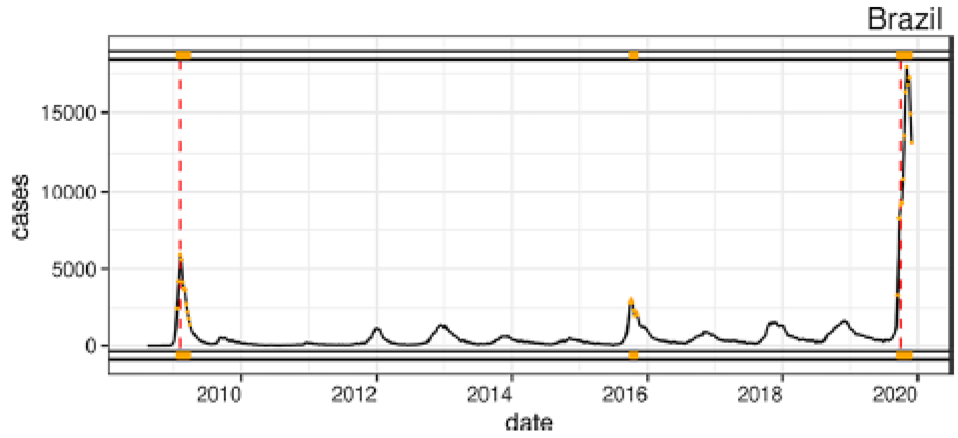
Keywords COVID-19 · Under-reporting · SARI · Time series modeling · Data analytics

Lais Baroni
lais.baroni@eic.cefet-rj.br

[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

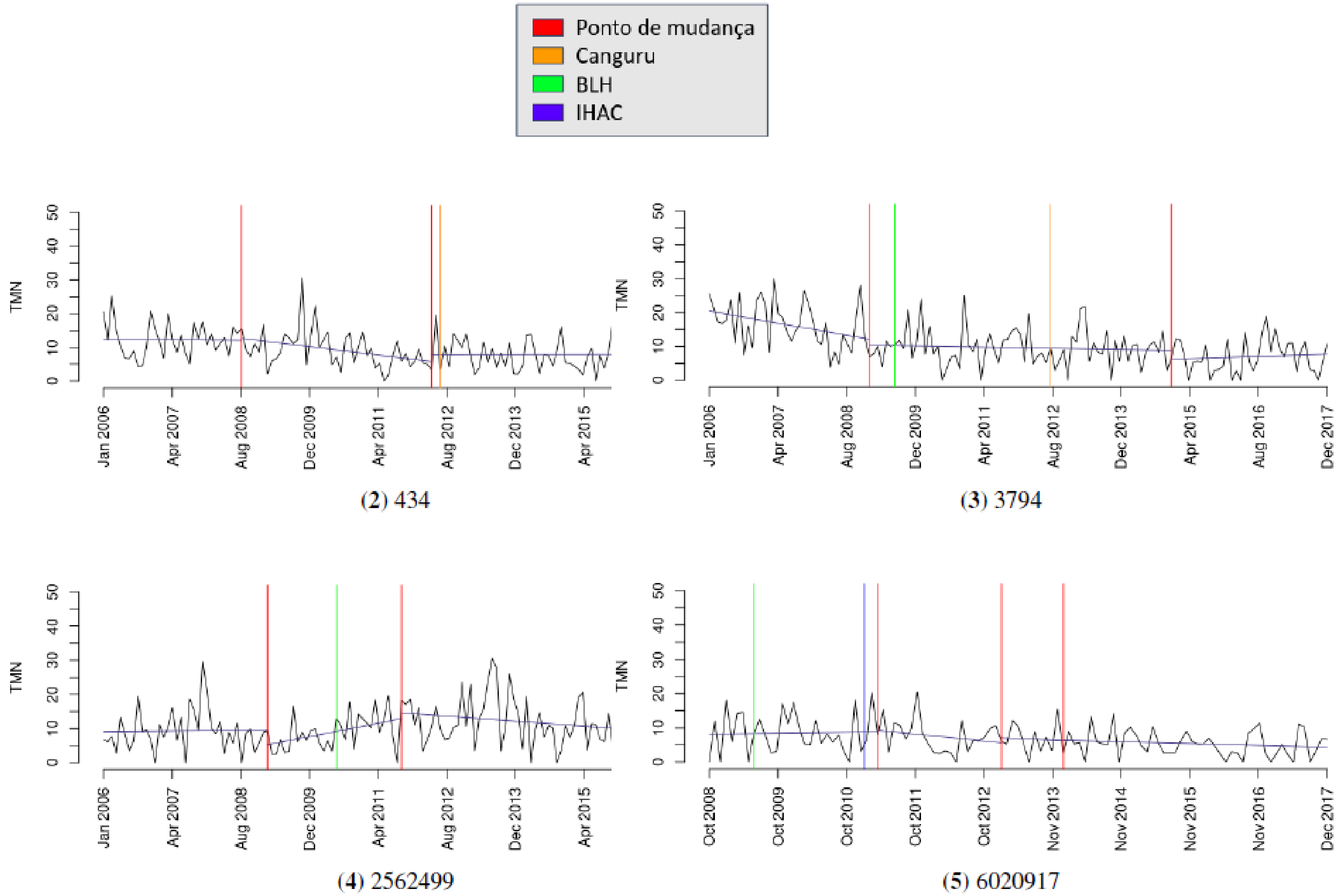
Published online: 14 March 2021

- Event detection



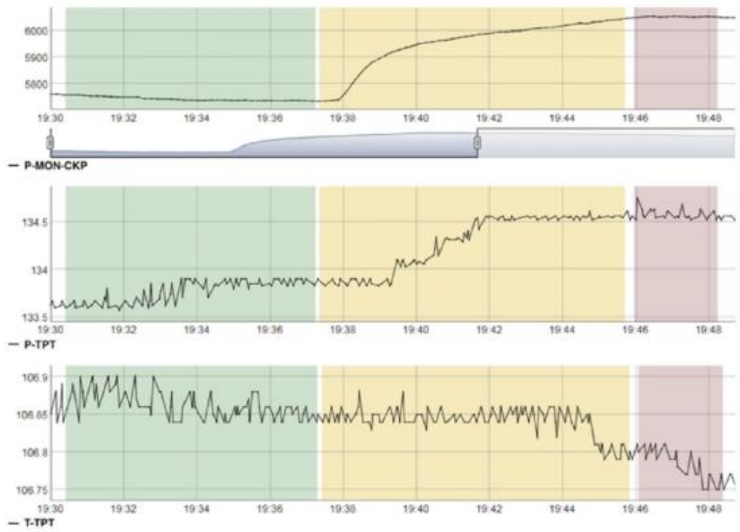
35

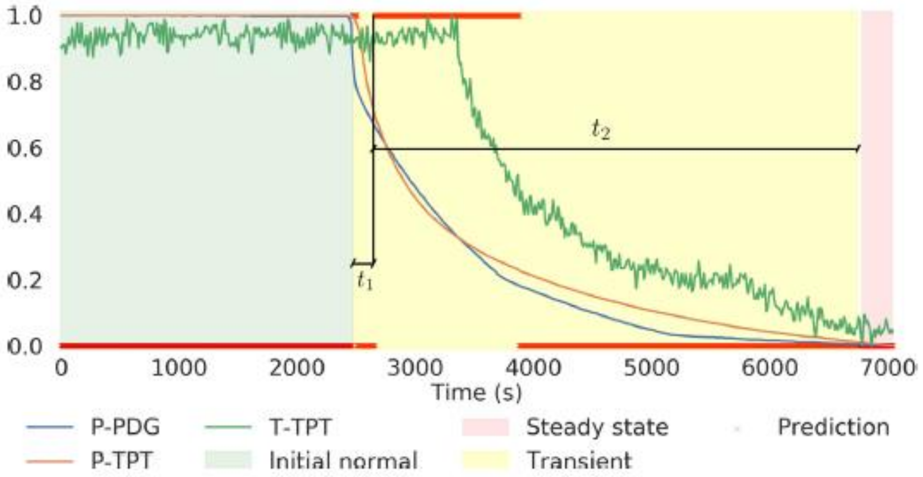
Applications: Event detection for neonatal mortality



Applications: Event detection for oil well drilling and exploration

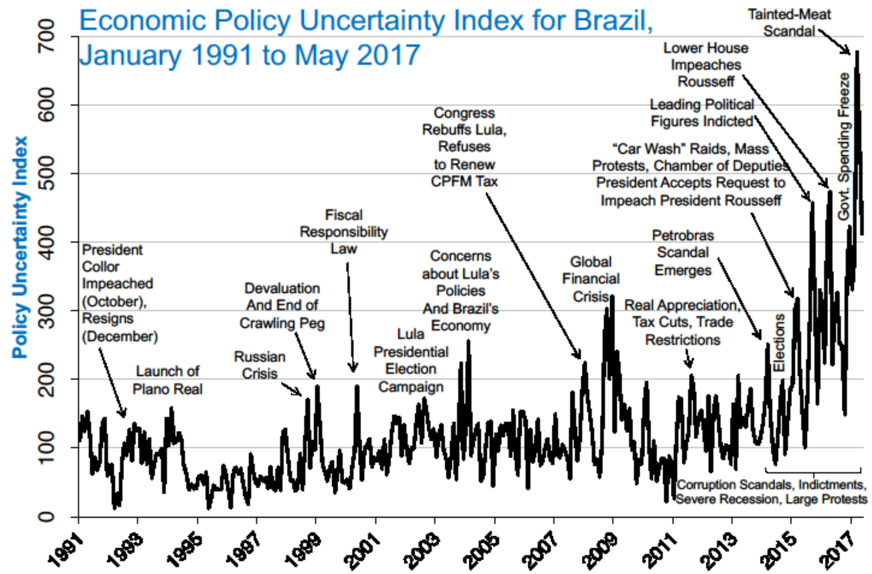
- Well drilling: Stuck pipe dataset
- Oil exploration: 3W dataset*





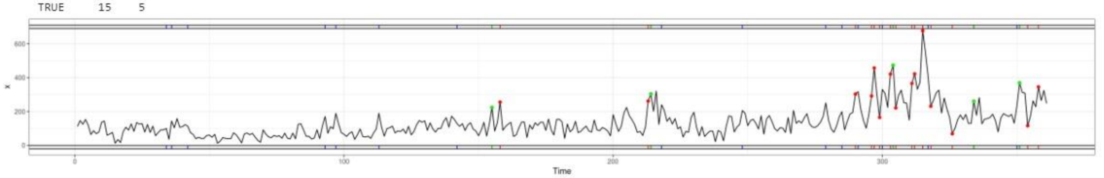
[1] R. E. V. Vargas et al. A realistic and public dataset with rare undesirable real events in oil wells. Em: Journal of Petroleum Science and Engineering 181 (2019), p. 106223.ISSN: 0920-4105.
[2] Marins, M. A., Barros, B. D., Santos, I. H., Barrionuevo, D. C., Vargas, R. E., Prego, T. D. M., ... & Netto, S. L. (2020). Fault detection and classification in oil wells and production/service lines using random forest. Journal of Petroleum Science and Engineering, 107879.
*Resultados preliminares se encontram no 4º relatório do projeto

Applications: Finance data
Economic Policy Uncertainty



Notes: Index reflects scaled monthly counts of articles in Folha de São Paulo containing "incerto" or "incerteza", "econômico" or "economia", and one or more policy-relevant terms that include regulação, déficit, orçamento, imposto, "banco central", planalto, congresso, senado, legislação, and tarifa. Normalized to a mean of 100 from 1991 to 2011. Index methods follow "Measuring Economic Policy Uncertainty" by Baker, Bloom and Davis. Data available at www.PolicyUncertainty.com.

```
           Reference
Prediction FALSE TRUE
     FALSE   326   15
     TRUE     15    5
```



[1] R. E. V. Vargas et al. A realistic and public dataset with rare undesirable real events in oil wells. Em: Journal of Petroleum Science and Engineering 181 (2019), p. 106223.ISSN: 0920-4105.
[2] Marins, M. A., Barros, B. D., Santos, I. H., Barrionuevo, D. C., Vargas, R. E., Prego, T. D. M., ... & Netto, S. L. (2020). Fault detection and classification in oil wells and production/service lines using random forest. Journal of Petroleum Science and Engineering, 107879.

*Resultados preliminares se encontram no 4º relatório do projeto

Event detection team

Cristiane Gea

Lais Baroni

Luciana Escobar

Rebecca Salles

Balthazar Paixão

Heraldo Borges

Janio Lima

Leonardo Carvalho

Lucas Giusti