INDRA-ASTRA

Online Multiscale Method for Automated Data-Quality Monitoring

Ronglong Fang [1]

in collaboration with Markus Diefenthaler[2] Abdullah Farhat [1] Holly Szumila-Vance[2] Yuesheng Xu [1]

[1]Old dominion university, Norfolk, VA, USA.

[2]Jefferson lab, Newport news, VA, USA

CONFERENCE ON COMPUTING IN HIGH ENERGY NUCLEAR PHYSICS, May 8, 2023

Content

Automated Data-Quality Monitoring

Online automated data-quality monitoring

1. Increase reliability of the data

   In data science, the results of data analysis directly depend on the quality of the data. Monitoring data online can improve the reliability of data.

2. Find and fix issues in near real-time.

   Online data-quality monitoring can find problems in data while data taking. In experimental physics, it is possible to take a month or longer to obtain data. By monitoring the data, we can improve the stability of the detector by detecting issues in the data

Overview of the problem

To deal with online data, our goal is :

1. detects when a change occurs.
2. determines what kind of change occurs, e.g., a sudden change occurs or linear gradual change occurs.
3. Autonomous calibration using basedline calibrations.

Multiscale method

Our approach

1. Using multiscale basis to represent the data and the coefficient of basis store the information for the data.
   (a) If no change happens, the amplitude of the coefficient of the multiscale basis is small.
   (b) If a change happens, the amplitude of the coefficient of the multiscale basis is big.

   change in raw data set $\rightarrow$ outlier in coefficients set

2. Detect outlier in the coefficients set. The outlier coefficients indicate that changes happen in its support.

Multiscale representation for the raw data

For a fixed scale parameter k and the data set

$$\left[d_j, d_{j+1}, \ldots, d_{j+2^k-1}\right],$$

the coefficient for this data is defined as

$$a_j^k = \frac{1}{2^{2k}} \sum_{i=0}^{2^k-1} d_{j+i}\psi_k(x_i), \tag{1}$$

where ψ_k is the k-th level basis, the $x_i := \frac{i}{2^{2k}} + \frac{0.5}{2^{2k}}$ are the corresponding discrete nodes, $\left[d_j, d_{j+1}, \ldots, d_{j+2^k-1}\right]$ is the support of the coefficient a_j^k.

The basis function ψ defined on $[0, 1]$ in (1) satisfies the following properties :

1. Vanishing moment [1] property of order n, that is

$$\int_0^1 \psi(x) x^j dx = 0, \ j = 0, 1, \ldots, n-1. \tag{2}$$

2. The nonzero part of ψ is a subset of $[0, 1]$.

The choice of basis

- piecewise constant test function

$$\psi(x) = \begin{cases} 1, & 0 \leqslant x \leqslant 1/2 \\ -1, & 1/2 < x \leqslant 1 \end{cases} \tag{3}$$

- piecewise linear test function

$$\psi(x) = \begin{cases} 1 - 4x, & x \in [0, \frac{1}{2}] \\ 4x - 3, & x \in (\frac{1}{2}, 1] \end{cases} \tag{4}$$

Online multiscale algorithm

Online multiscale algorithm

The structure of the online multiscale algorithm can be described as following

$$\underbrace{d_0, d_1}_{a_0^1}$$

Online multiscale algorithm

The structure of the online multiscale algorithm can be described as following

$$\overbrace{\underbrace{d_0, d_1}_{a_0^1}, \underbrace{d_2, d_3}_{a_2^1}}^{a_0^2}$$

Online multiscale algorithm

The structure of the online multiscale algorithm can be described as following

$$\overbrace{\underbrace{d_0, d_1}_{a_0^1}, \underbrace{d_2, d_3}_{a_2^1}}^{a_0^2}, \underbrace{d_4, d_5}_{a_4^1}$$

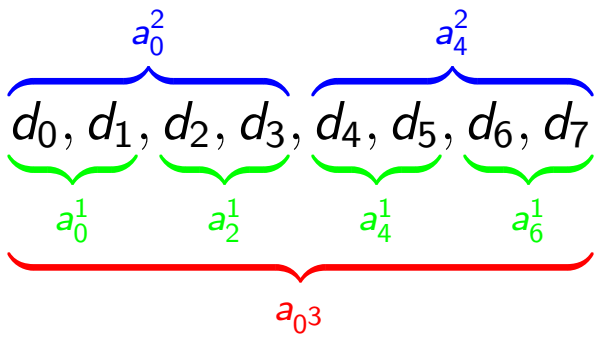
Online multiscale algorithm

The structure of the online multiscale algorithm can be described as following

$$\overbrace{\underbrace{d_0}_{a_0^1}, \underbrace{d_1}_{a_2^1}, \underbrace{d_2}_{a_4^1}, \underbrace{d_3}_{a_6^1}}^{a_0^2}, \overbrace{\underbrace{d_4}_{a_4^1}, \underbrace{d_5}_{a_6^1}, \underbrace{d_6}_{a_8^1}, \underbrace{d_7}_{a_{10}^1}}^{a_4^2}$$

Online multiscale algorithm

The structure of the online multiscale algorithm can be described as following



Online multiscale change detection algorithm

Algorithm 1: Online multiscale change detection and calibration

Data: sequential data $d_0, d_1, \cdots, d_i, \cdots$; test function ψ; the minimum scale k_{min} and the maximum scale k_{max}. W denotes the empty data set, $W :=$ {add the first $2^{k_{max}}$ data}; Let $a := \left\{a^{k_{min}}, a^{k_{min}+1}, \cdots, a^{k_{max}}\right\}$ denotes the coefficient for each scale and each element is an empty set; let CInter := $\{[0, 0]\}$ denotes the change interval.

Result: return the calibrated data.

$\ell = 2^{k_{max}-k_{min}}$,
for $j = \ell, \ell + 1, \cdots$ **do**
  $W := W \cup \{$ the latest $2^{k_{min}}$ data $\}$,
  $m = k_{min}$,
  while $m \leq k_{max}$ **do**
    if $floor\left(\frac{len(W)}{2^m}\right) - floor\left(\frac{j}{2^{m-k_{min}}}\right) > 0$ **then**
      let $s_0, s_1, \ldots, s_{2^m-1}$ to denote the latest 2^m new data, calculate scale m coefficients for this data set, denote as a_{new}. Store a_{new} to the m scale coefficients set,

$$a^m = a^m \cup \{a_{\text{new}}\}.$$

      detect whether a_{new} is an outlier in m scale coefficients set a^m.
      if a_{new} *is an outlier in the the set* a_m **then**
        receive a change interval, denote as $[l_j^m, r_j^m]$.
        calibrate the changed data.
      end
    end
  end
end

Outlier detection

We detected outliers based on sample mean and sample variance, which is defined as

$$\text{Out}_t := \{a : a \notin [\hat{\mu} - t\hat{\sigma}, \hat{\mu} + t\hat{\sigma}]\} \tag{5}$$

where $\hat{\mu}, \hat{\sigma}^2$ are sample mean and sample variance, and t is a predefined threshold.

Autonomous calibration using basedline calibrations

We use the baseline to calibrate,

$$d_{new}\left[l:r\right] := d\left[l:r\right] - \left(\mu + th \times \sigma\right), \tag{6}$$

where μ, σ be the mean and standard deviation of the data in the interval $[l, r]$, and th is the calibrated parameter.

The purpose of this calibration method is to remove the background of the data set.

Results for Physics Data

GEM data

1. The data is taken during the Jefferson Lab Hall A Super Big Bite (SBS) experiments. The SBS experiments are characterized as high rate counting experiments and employ new Gaseous Electron Multiplier (GEM) detectors for tracking.

2. Gaseous Electron Multiplier (GEM) data is used to reconstruct the track of particle and then to infer information about the particle's origin and momentum.

3. The size of data is 30006912.

4. We add sudden change and linear gradual change artificially. The goal is to detect the change in the data set.

Original Data and changed data

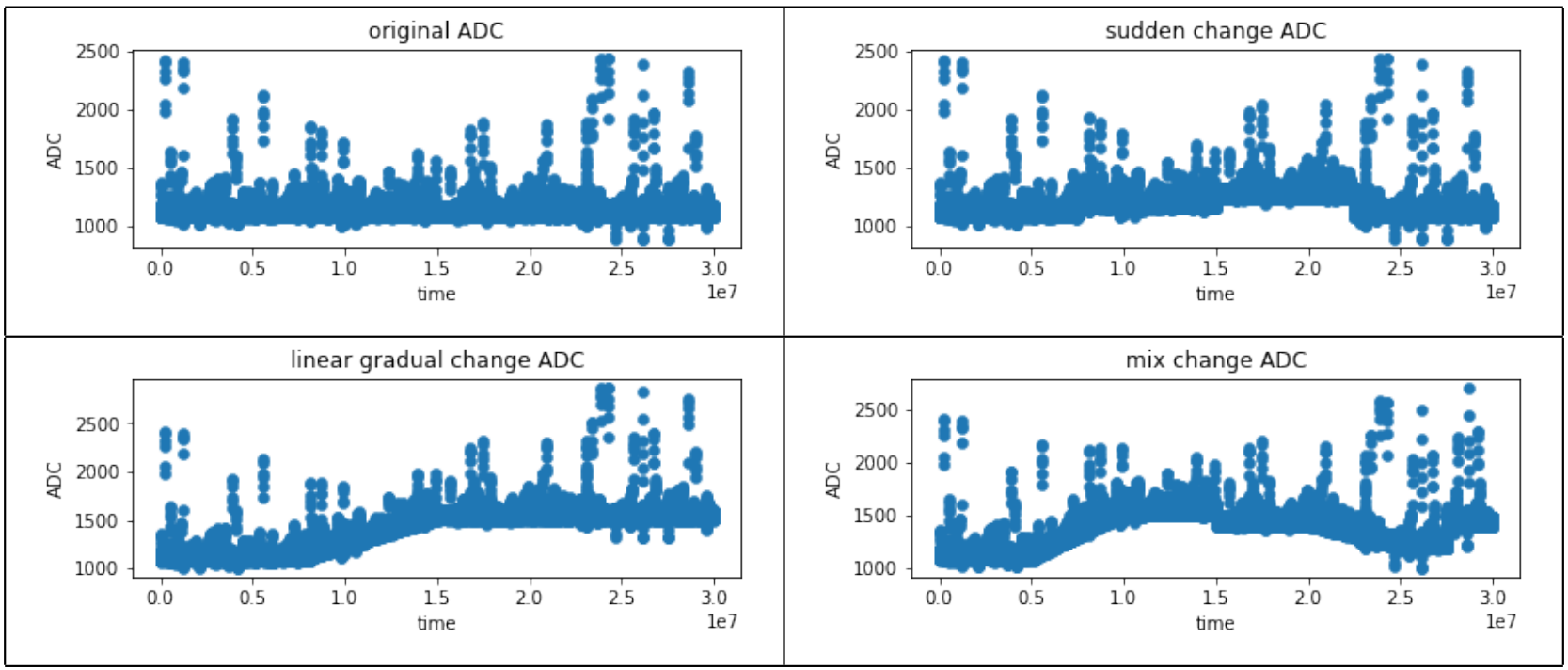


Figure 1: Raw data and changed data

Sudden change

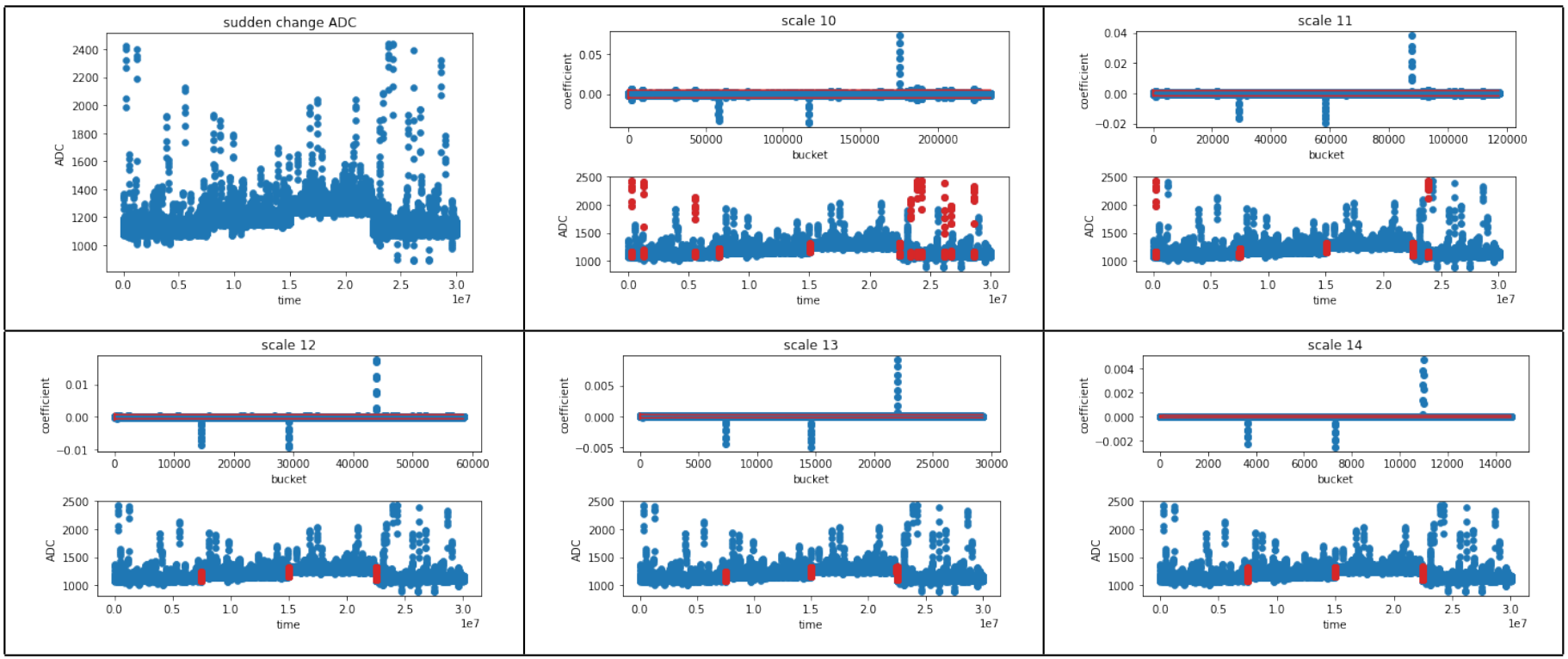


Figure 2: Sudden change results : scale 10-14, threshold 10.

The multiscale representation magnifies the sudden change and shrinks the noise in the raw data set. When we increase the scale, the results is more reliable, however, the accuracy goes down.

Gradual change results

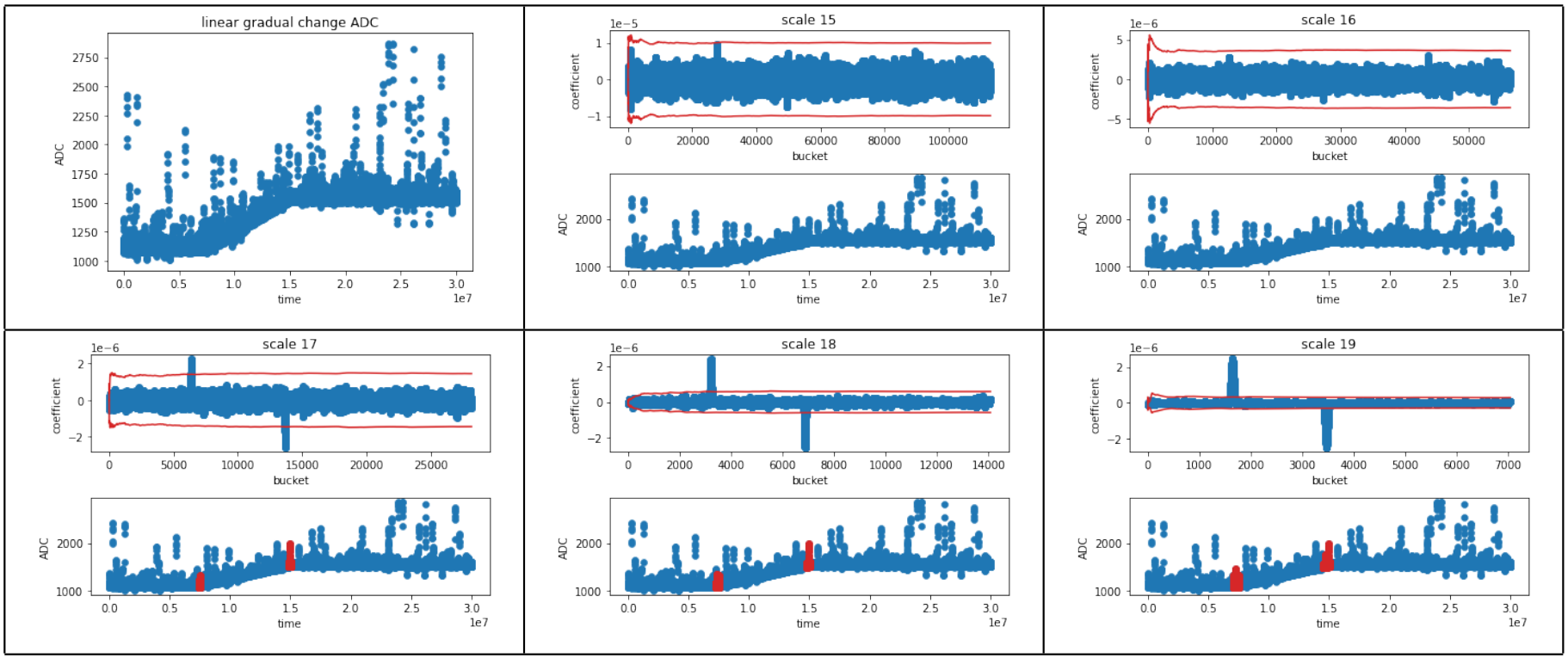
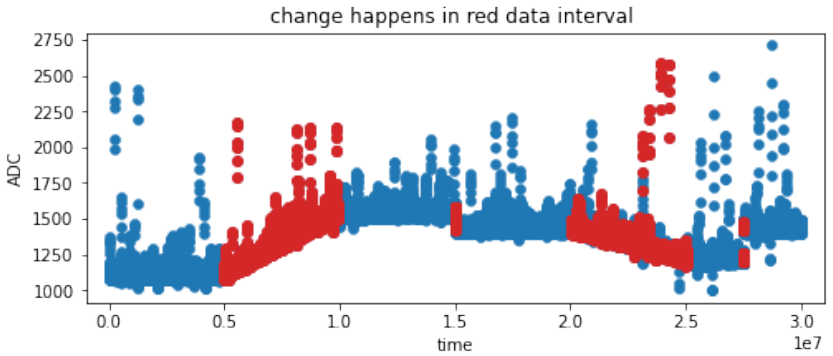


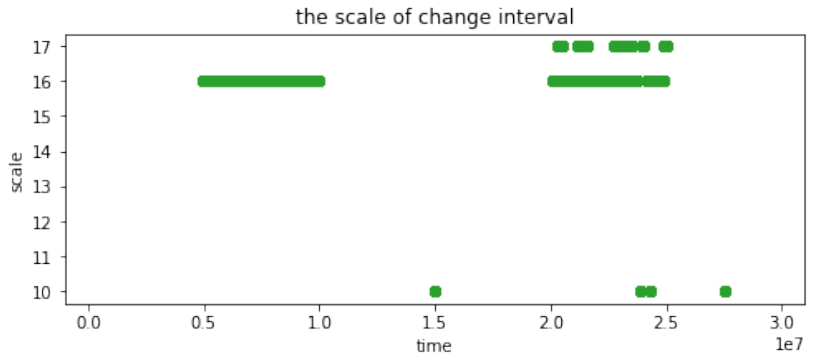
Figure 3: linear gradual change results : scale 15-19

For the linear gradual change, we got a similar results. The gradual change is been detected in a higher level.

Mix change



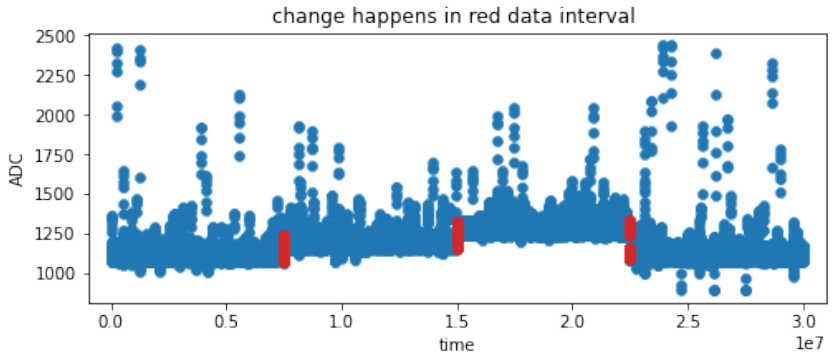
(a)



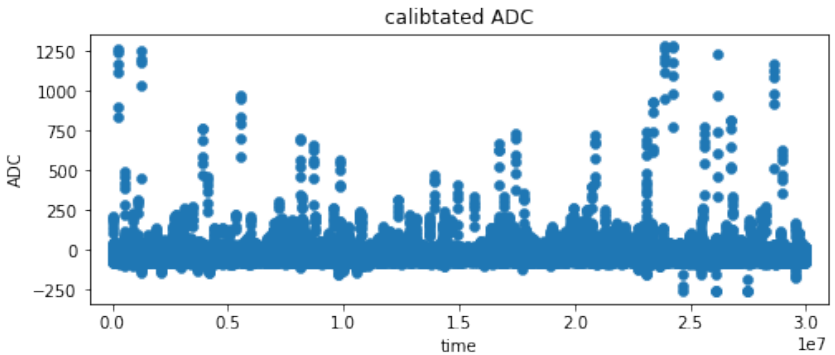
(b)

Figure 4: The minimal and maximal scale we set is 10 and 20. (a) shows the results of mix change detection results when using a piecewise constant basis and choosing an outlier parameter $t = 11$ and (b) is the corresponding scale for each change interval.

Sudden change calibration



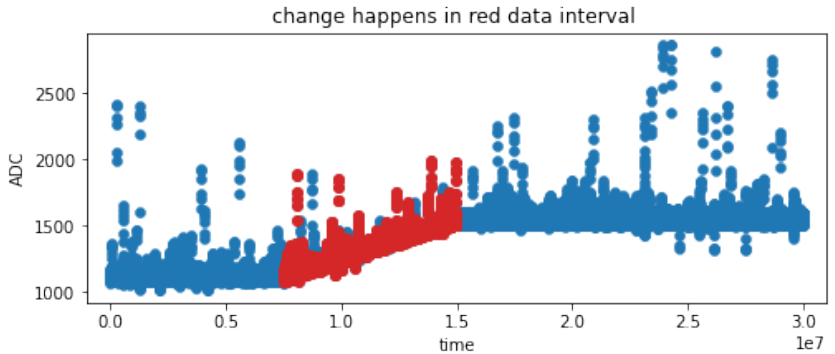
(a)



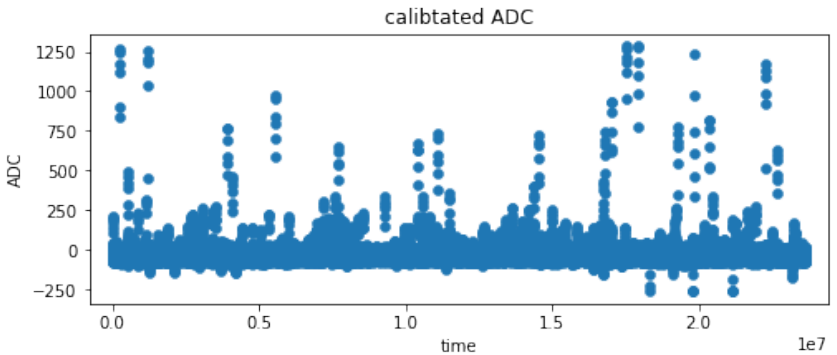
(b)

Figure 5: (a) shows the results of sudden change detection results when using a piecewise constant basis and choosing an outlier parameter $t = 11$ and (b) is the corresponding calibrated ADC based on the detection results .

Linear graduate change calibration



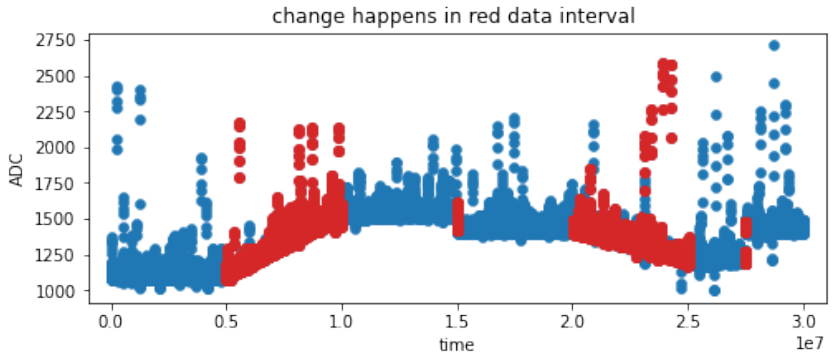
(a)



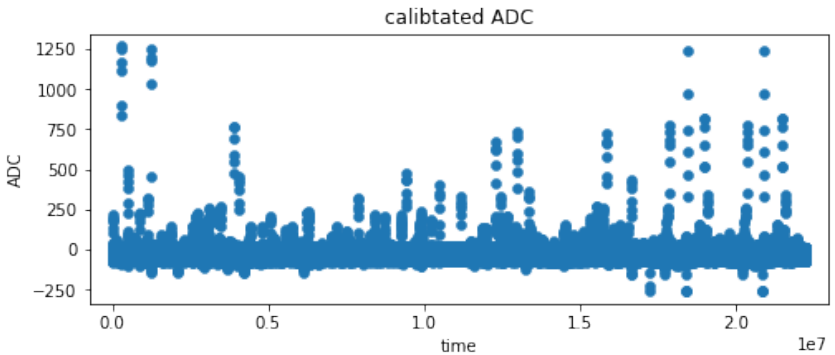
(b)

Figure 6: (a) shows the results of linear change detection results when using a piecewise constant basis and choosing an outlier parameter $t = 11$ and (b) is the corresponding calibrated ADC based on the detection results .

Mix change calibration



(a)



(b)

Figure 7: (a) shows the results of mix change detection results when using a piecewise constant basis and choosing an outlier parameter $t = 12$ and (b) is the corresponding calibrated ADC based on the detection results .

Summary

1. Using multiscale basis to represent the raw data set.

   $$\text{change in raw data set} \rightarrow \text{outlier in coefficient set}$$

   Develop an online multiscale method to monitor the data.

2. Autonomous calibrations using autonomous change detection and baseline calibrations.

3. The online multiscale method is an alternative to machine learning and doesn't require any training.

Thank you


Z. Chen, C. A. Micchelli, and Y. Xu, *Multiscale methods for Fredholm integral equations*, vol. 28, Cambridge University Press, 2015.